

Electricity Cash Out Arrangements

Stephen Littlechild¹
9 March 2007

The Office of Gas and Electricity Markets (Ofgem) is carrying out a review of electricity cash out arrangements – that is, the arrangements whereby generators and suppliers pay or are paid for imbalances (shortages and surpluses of power relative to their contracted commitments). Ofgem has asked me to report on what electricity cash out arrangements for the Great Britain (GB) market should look like. The Report should cover elements of cash out such as the derivation of imbalance price(s), the role of the system operator, incentives on market participants, how costs are targeted, and any other relevant aspects. The Report may point to practical obstacles, but overall should focus on principles. If time allows, it should draw on international comparisons. Appendix 1 sets out the terms of reference in full.

The content of the Report is as follows:

- Section 1 explores the purpose or aims of cash out arrangements.
- Section 2 considers whether cash out should be seen as a market or as a service, and whether it is more helpful to aim at a balancing market (as illustrated by the Electricity Reliability Council of Texas (ERCOT) arrangements) or at a balancing mechanism that gives greater discretion to the System Operator and provides cash out services other than via a market.
- Section 3 examines the case for dual cash out versus a single cash out price.
- Section 4 discusses the issue of average versus marginal cost pricing.
- Section 5 examines cost allocation and tagging.
- Section 6 looks at ex ante pricing and ex post trading.
- Section 7 discusses innovation and markets, governance and further research.
- Section 8 provides a summary of conclusions.

Earlier documents have summarised and discussed the history and nature of cash out arrangements to date.² I have not tried to document all the views to which I refer here, though for convenience I have cross-referenced some of the points in Nigel Cornwall's parallel paper for Ofgem.³ However, I have not sought to grapple with the many operational issues that he covers. I hope the two papers complement each other in their analyses and recommendations, and promote fruitful discussion.

¹ Emeritus Professor, University of Birmingham, and Senior Research Associate, Judge Business School, University of Cambridge. As an independent consultant on privatisation, competition and regulation over the past eight years I have advised numerous companies and regulatory organisations in the electricity sector. I have recently been invited to be a non-executive director of a company that in due course will seek to be listed on the AIM and operate as an electricity supplier.

² See for example *Electricity and gas cash out review: a consultation document*, Ofgem, May 2004.

³ Nigel Cornwall, *Cash-out Review 2007: "An Independent Perspective"*, March 2007. (Henceforth cited as Cornwall, *Cash-out Review 2007*)

I have had limited previous involvement with cash out arrangements.⁴ I have not followed in detail the day-to-day operation and various revisions of the cash out arrangements, but have occasionally commented on some aspects thereof.⁵ In the time available I have not sought to master the extensive literature on these topics or to canvass opinion among market participants generally.⁶ I have had helpful clarification on various issues, including from Ofgem, National Grid, Nigel Cornwall and Dr Parviz Adib, Director of Wholesale Market Oversight (WMO) at the Public Utility Commission of Texas (PUCT). None of these is to be held responsible for the content or conclusions of this Report.

⁴ As Director General of Electricity Supply (DGES) from 1989 to December 1998, I recommended changing the market in England and Wales from the Pool to what became the New Electricity Trading Arrangements (NETA). See *Review of Electricity Trading Arrangements: Proposals*, Office of Electricity Regulation (Offer), July 1998. However, I was not involved in the subsequent detailed design of NETA arrangements, such as the nature of the cash out system, nor in the implementation of NETA in March 2001 or its extension to Scotland under the British Electricity Trading and Transmission Arrangements (BETTA) in April 2005.

⁵ See for example *Electricity: Regulatory Developments Around the World* (Beesley Lecture, London, 9 October 2001) in C Robinson (ed) *Competition and Regulation in Utility Markets*, London: Institute of Economic Affairs and London Business School, 2003, pp 61-87; and section 9.3 of my Report for Ofgem *Smaller Suppliers in the UK Domestic Electricity Market: Experience, Concerns and Policy Recommendations*, 29 June 2005 (published 5 July 2005), available at <http://www.electricitypolicy.org.uk/pubs/misc.html>.

⁶ It is nonetheless worth noting one of the most recent and thorough reviews of the British electricity market, which has several sections on the cash out arrangements. David Newbery, *Electricity Liberalization in Britain and the Evolution of Market Design*, in Fereidoon P Sioshansi and Wolfgang Pfaffenburger (eds) *Electricity Market Reform: An International Perspective*, Elsevier, 2006.

1. The purpose of cash out arrangements

1.1 The transition from the Pool to NETA

In order to understand the nature and aims of the present electricity market (NETA and subsequently BETTA) it is helpful to look briefly at the Pool arrangement that preceded it, and the reasons for the change in approach.

The Electricity Pool of England and Wales, and the associated arrangements for contracts for differences, were an ingenious and in many respects remarkably successful way of modifying the arrangements for running the nationalised monopoly that was the Electricity Supply Industry. These arrangements enabled privatisation to take place and competition to be introduced. The Pool provided a public price for electricity for each half hour, on a day-ahead basis.

However, the Pool also had limitations. In simple terms, the Pool was only half a market: generators competed to supply into the Pool but they were not faced by buyers competing to take from the Pool. Prices were not determined by agreements between willing buyers and sellers, as in any normal competitive market. Instead, the Pool was essentially a single buyer model. With minor exceptions, all electricity had to be sold to the Pool, and all electricity had to be bought from the Pool. The System Operator (as it was commonly called in other jurisdictions) made the decisions about what plant to run and when, and it set the rules for determining prices. It did so based on specified procedures and an approved scheduling model that was a modified version of that used by the Central Electricity Generating Board (CEGB).

It is true that there was competition among generators to be selected to run. And that financial contracts for differences enabled buyers and sellers to agree contract prices in place of the Pool price. Some said that the Pool bidding rules plus contracts made it possible to do anything a market participant would want to do. Nonetheless, this was a far cry from any other competitive industry, in which buyers and sellers negotiated and agreed terms on a bilateral basis, subject to competition from other market participants but not subject to such extensive control by an intermediate authority.

There was therefore a case for going beyond the Pool. There was a need to complete the transition from the largely internal scheduling arrangements appropriate to the CEGB as a nationalised monopoly to the trading arrangements appropriate to a competitive market. The Pool was the first step in this process but not the last. NETA was thus an important part of the transition to a more competitive electricity market.

1.2 Facilitating a competitive electricity market

NETA was intended to give maximum flexibility to market participants to make their own arrangements for buying and selling electricity. These might be bilaterally negotiated contracts ranging from a few hours duration to many years duration, and

signed an hour ahead or years ahead. There might be standard contracts traded in a more formal market place. It might mean own generation in the case of a vertically integrated company.

Henceforth, the role of the System Operator, put crudely, was to take orders instead of to give orders. That is, the System Operator would receive the plans of the market participants and as far as possible enable them to be carried out. Instead of determining the dispatch of generating plant, the System Operator would be informed of the bilateral and other physical contracts entered into by generators and suppliers at different nodes, and seek to implement them consistent with the capacity of the transmission system and with stability of the system.

Of course, certain ground rules had to be established. Suppliers had to ensure that they arranged to put into the system as much power as they took out, including to cover network losses. The parties had to pay sufficient charges to cover the reasonable costs of an efficient System Operator.

From this perspective, the specific function of cash out arrangements was two-fold. In the first place, it would enable market participants to buy and sell electricity even though they were not, and could not be, 100 per cent sure what the precise level of their customer demand or generator output would be at any particular time. The cash out arrangements would ensure that any unintended shortages or surpluses by individual market participants would be made good or disposed of. Market participants would be charged or reimbursed appropriately, to reflect the costs involved.

At the same time, cash out arrangements would enable the System Operator to arrange, in an efficient way, the increases or decreases in generation necessary to maintain system stability. In doing so it would make good the unintended shortages and surpluses, or imbalances, experienced by individual market participants.

The underlying objective is thus to facilitate the effective and efficient operation of a competitive electricity market. The specific interpretation and implementation of this objective has of course evolved over time. Nevertheless, the underlying objective remains valid.

1.3 Further implications

At least three aspects or implications of the underlying objective deserve further explicit acknowledgement.

First, it implies facilitating competition in generation and supply rather than facilitating the exercise of market power. Where possible, the arrangements should be conducive to new entry and not set up artificial barriers. They should not bear unduly heavily on any particular class of market participants, particularly new entrants.

Second, a competitive electricity market is desirable not just for its own sake, but because it will also be conducive to improved efficiency in the sector as a whole. For example, it will encourage market participants to discover the goods and services that customers want, and to produce those goods and services at lowest cost, with continual innovation. It will also encourage movement towards an efficient structure of the industry. For example, relevant questions are how far vertical integration is more efficient than vertical contracting, what kinds and durations and terms of contracts are most efficient, what kinds of market places are needed to enable generators and suppliers to transact. The introduction of NETA meant that all these are for the market to discover, rather than for the government or the regulator or the System Operator (and associated voters in the Pool) to determine. Cash out arrangements should therefore be consistent with this.

Third, looking specifically to the present topic, the cash out arrangements should be conducive to an efficient pattern of balancing within the market as a whole. That is, consistent with maintaining system stability, they should seek to ensure that market participants balance their own positions where and to the extent that this is the most efficient way of doing things, and should use the facilities provided by the system operator when and where this is a more efficient way to do things. Amongst other things this means seeking to ensure that cash out charges reflect the costs of providing cash out services, so that market participants are incentivised to take market-based decisions that are consistent with the efficiency of the market as a whole.

1.4 Market power

A concern in the early days was that the Pool was conducive or vulnerable to the exploitation of market power. It was hoped that NETA with its focus on bilateral trading would be less vulnerable to the abuse of market power. The design of the cash out arrangements perhaps reflected this. For example, if dominant generators had been able to manipulate Pool prices based on bidding arrangements with marginal pricing, there might be a danger that they could similarly manipulate cash out prices if calculated on a similar basis. If a large volume of electricity were to pass through the cash out mechanism instead of via bilateral trading, this could reintroduce the same problems of market power abuse. A dual cash out rather than single cash out price might reduce the likelihood of this happening to the extent that it would incentivise market participants to enter into contracts to balance their own positions before gate closure, rather than to rely unduly on the cash out mechanism.

Since that time, generator market power has been reduced, at least in the sense that there are now half a dozen major players, as well as other substantial generators, rather than a duopoly owning most of the price setting plants. Also, recent econometric work has cast doubt on how far the subsequent observed reductions in wholesale market prices were in fact due to NETA as opposed to restructuring of the generating sector.⁷

⁷ Evans, J.E. and R.J. Green, *Why Did British Electricity Prices Fall After 1998?* Department of Economics Working Paper 05-13, University of Birmingham, 2005; also MIT Center for Energy and Environmental Policy Research WP 2003-007. I am told that there is also some theoretical literature suggesting that net Pools (allowing bilateral physical trading) may have similar outcomes to gross Pools, but I have not explored this.

This suggests that the future development of cash out arrangements need not be so concerned about tackling generator market power, if indeed they are capable of affecting that. Cash out arrangements should nonetheless still be designed where possible to reduce barriers to entry into generation and supply and to facilitate competition in the electricity sector generally. Some would suggest that the complexity of the present rules is not fully conducive to this.

1.5 Facilitating the task of the System Operator

Another consideration in design of initial cash out arrangements may have been the implications for the tasks and risks faced by the System Operator in balancing the system. Requiring the System Operator to accommodate a set of physical bilateral contracts was quite different from allowing it to schedule and dispatch generation itself. Other things being equal, it would be easier for the System Operator to discharge its new duties in a system that was more nearly in balance as a result of the decisions and plans of market participants, than in a system that was far from being balanced. This would be a particularly important consideration at the time when new arrangements were being introduced, when the policies of market participants were as yet quite unknown, and when the effects of actions taken by the System Operator were less well understood. In such circumstances, it was more difficult to judge the ability of the System Operator to meet the system demands in the time available and the extent and consequences of any difficulties that might be encountered.

From this perspective, prudence was called for. It was considered helpful to have a set of cash out arrangements that actively encouraged parties to plan to be in balance. To that end, cash out arrangements that penalised imbalances – or at the very least provided strong incentives to avoid imbalances – were considered appropriate.⁸

In the event, the NETA arrangements generally worked well and the System Operator coped successfully with balancing the system. Indeed, it soon became possible to advance gate closure from three and a half hours to one hour. Other systems are nowadays working almost in real time. This suggests that, while the System Operator's ability to ensure that the system stays in balance is of course still important, it is no longer the critical consideration that it once was thought to be. The implications for the System Operator need to be considered, but this should not dominate or unduly constrain the future design and development of cash out arrangements.

⁸ “The intention is to provide *stronger incentives than at present* for generators and suppliers to meet their commitments”. *Review of electricity trading arrangements: Framework Document*, Offer, November 1998 (para 2.3, italics added). This was appropriate since there were little or no such incentives under the Pool. I have not tried to document whether, when or to what extent subsequent arrangements embodied incentives going beyond cost-reflectivity.

1.6 Conclusion on the purpose of cash out arrangements

The design and implementation of NETA and the cash out arrangements reflected a general underlying objective of facilitating a competitive and efficient electricity market. At the time there were also concerns about market power and the ability of the System Operator to cope with the new approach. It is perhaps not surprising if statements from Offer and later Ofgem reflected this multi-faceted situation, and did not focus specifically on a single aim or objective. Some six years later, however, the two main concerns – about generator market power and about the ability of the System Operator to balance the system - no longer seem so serious or risky as they were in the early years.

The underlying objective of facilitating the operation of a competitive and efficient electricity market remains as valid as ever. Arguably, this can now be the prime determinant of how the cash out arrangements are designed. Those features that reflect the earlier concerns can now be modified if they are not conducive to this prime objective, or if they impact unduly or adversely on particular types of competitors or potential entrants.

2. Cash out: market or service?

Many electricity jurisdictions describe themselves as having a ‘balancing market’, whereas in GB the phrase ‘balancing mechanism’ is used. Is there any significance in this distinction? Does it mean, for example, that GB prefers a non-market means of providing balancing services? Or are they simply different terms for the same thing? If there is a difference, what are the pros and cons of each approach? Is there a case for making the GB balancing mechanism more like the balancing markets found elsewhere? Or is there a case for further developing the concept of the balancing mechanism even though it may become even more different from a balancing market?

2.1 The balancing market in Texas

Several European electricity markets are based on bilateral trading and have rules comparable to those in NETA and BETTA. However, they are less competitive in objective and in practice than the GB market. A more useful comparative example of a balancing market is the ERCOT system in Texas.⁹ This began to operate as a single control area market on July 31 2001. (Prior to that, generation dispatch decisions and other operational decisions were made locally in ten control areas.) Internationally, Texas is perhaps the most similar market to GB, in the sense that the ERCOT wholesale electricity market is based primarily on bilateral contracts rather than on a centrally dispatched bid-based Pool. Balancing trades account for some 3 to 5 per cent of end-user requirements, as is the case in GB. Texas is also a relatively competitive electricity market. Retail competition is more active there than in any other US system, to the extent that retail price controls (the ‘price to beat’) were removed in January 2007.

A variety of formal and informal markets have developed in ERCOT. These include the bilateral and day-ahead markets (not operated by ERCOT); the Balancing Energy Services (BES) and Ancillary Services including day-ahead operating reserves markets (administered by ERCOT); Transmission Congestion Rights (auctioned by ERCOT); and Renewable Energy Credits (RECs) (administered by ERCOT).

Market participants are required to submit their schedules of energy, matching generation to load, to the ERCOT Independent System Operator (ERCOT ISO) through a Qualified Scheduling Entity (QSE). For every 15-minute settlement interval the ERCOT ISO compares the sum of the schedules submitted by QSEs to its own load forecasts and determines balancing energy and Ancillary Services requirements. Similarly, for every operating hour the ERCOT ISO also determines Ancillary Services requirements in day-ahead markets.

⁹ The following account is based primarily on R Baldick and H Niu, Lessons learned: the Texas experience, in James Griffin and Steven Puller (eds) *Electricity Deregulation: Where to from here?*, University of Chicago Press, 2005, available at <http://www.ece.utexas.edu/~Baldick/papers/papers.html>, and Parviz Adib and Jay Zarnikau, Texas: the most robust competitive market in North America, in Fereidoon P Sioshansi and Wolfgang Pfaffenburger (eds) *Electricity Market Reform: An International Perspective*, Elsevier, 2006. I am grateful for further clarification from the WMO of the PUCT.

Initially the individual schedules of energy were required to be balanced, as a way of minimizing balancing energy volumes, which would create less credit and financial risk for the ERCOT ISO. However, some market participants were concerned that this restricted their ability to buy and sell energy actively so as more effectively to manage their portfolio risks. Industrial and large Commercial loads wanted the ability to go short and to rely on their Retail Electric Providers (REPs) to purchase the remainder on their behalf from a spot market, or alternatively to curtail their demand.

In late 2002 the 'relaxed balanced schedule' was introduced ERCOT-wide, whereby QSEs were no longer expected to schedule demand equal to their forecast. Rather, they were required to be more accurate in reflecting their supply sources in their balanced schedule, and to rely on the balancing market for any imbalances from their total demand. The aim was to remove any disincentive to use the balancing market and to provide greater flexibility for market participants to decide the appropriate extent of activity in the balancing market, given their particular risk profiles and preferences. In 2003 the relaxed balanced schedule was modified to limit the allowable deviation between the schedules and forecast (subject to particular credit requirements to be calculated for each market participant).

The expectation was that a larger amount of energy would be transacted in the balancing market, increasing the liquidity in this market. In the event there was not a great increase in the total volume transacted in the BES, but some market participants now effectively relied on this option.

The market operations process contains three major periods:

- day ahead Ancillary Services market (6am to 6pm on day prior to operating day),
- adjustment period (from close of day ahead market to one hour prior to operating hour), and
- operating period (including the hour before).

In each of these periods, market participants submit bids and offers. The ERCOT ISO accepts those that it judges necessary to meet its needs, given the bids and offers submitted and the generation operating constraints. It then determines prices.

Energy settlements are based on 15-minute intervals. For each 15-minute interval, the Market Clearing Price of Energy (MCPE) is determined and announced ten minutes before the beginning of that period. Individual shortages and surpluses of energy are charged or paid on the basis of the relevant MCPE.

2.2 Market prices?

Balancing prices (MCPEs) in each of these 15-minute intervals are market prices in the same (limited) sense as were prices in the England and Wales Pool. They are based on competing bids and offers submitted to a single buyer/seller. But they are not the result of bilaterally agreed deals between multiple buyers and sellers. More precisely

- they are based on separate schedules of bids and offers, where a single entity (the ERCOT ISO) determines the quantities to be taken

- the ISO determines prices based on these bids and offers, using a scheduling algorithm that is a security-constrained optimisation process, and
- the determined prices are those of the 'next' MWh in the stack rather than the 'last' (i.e. highest price) MWh actually selected.¹⁰

The nature of this scheduling algorithm – called the Market Clearing Price Engine or Scheduling, Pricing and Dispatch (SPD) - has been widely discussed for some time. That is not to say that all market participants fully understand it or how it is applied in actual cases. The larger more experienced companies are believed to be able to replicate the process, but the smaller less experienced ones regard it as a 'black box'. Nevertheless, the ability to see a firm price ahead of each operating period reduces uncertainty for market participants, and is considered a benefit.

2.3 Differences between Texas and GB

Balancing arrangements in GB involve complex tagging and other rules in order to determine cash out prices. Even so, they are arguably less algorithmic and model-based than those in Texas. Yet systems such as Texas choose to regard the balancing arrangements as a 'market', while GB describes them as a 'mechanism'. There are obviously similarities, but the differences go beyond nomenclature in at least three respects.

First, there is a difference in the scope of powers and discretion granted to the System Operator. In the Texas ERCOT system, there are more stringent restrictions on what the ERCOT ISO can do. On a day ahead basis it can instruct certain units of out of merit capacity (OOMC) to be sure that the right units are on in the right locations to prevent local congestion and secure the reliability of the system. But the stakeholders made a decision in 2000 to prevent ERCOT from competing with them. For example, generators were not allowed to use units contracted under the Reliability Must Run (RMR) contracts to compete in the energy market. They had to return ninety per cent of the associated revenues to the ISO if such units provided energy in the BES market.

In contrast, the System Operator in GB has more discretion to purchase energy in advance if it deems this likely to reduce the costs of balancing the system, subject to not being allowed to speculate. In addition, though this is not a necessary concomitant, there seems to be less transparency about the actions that the System Operator takes in GB than about those that the ISO takes in Texas.

A second difference relates to pricing of imbalances. In Texas, all units in each balancing market period are paid the same price, nominally a market-clearing price. In effect, the ISO seems to be charged with the duty of facilitating a balancing market, or as close to the concept of a market as circumstances allow, and with minimal infringement on the workings of other electricity markets and on the operations of market participants. The

¹⁰ The resulting prices are sometimes referred to as 'shadow' prices. Since offers are submitted in blocks (usually 25MW), there is generally no difference between the 'last' (or 'marginal') and the 'next' (or 'shadow') price, except for occasions when the next unit comes from the next block in the bid stack.

assumption is that an efficient market implies a single market clearing price (also called a uniform price), set equal to the cost or bid of the marginal unit.

In contrast, the GB balancing mechanism works on the basis of ‘pay-as-bid’. The System Operator is seen as providing a service to market participants that goes beyond that of facilitating a balancing market. The System Operator is expected and incentivised to get the best deal for market participants by carrying out a more active role. As noted, this may be by purchasing power ahead of time in order to reduce the costs of balancing. In addition, insofar as the System Operator purchases power in the balancing mechanism, pay-as-bid was assumed to be a cheaper way to secure that power, by virtue of the System Operator’s monopsony position as a single buyer. (Some argue against this, as discussed below.)

Third, in Texas there is a single balancing price that market participants either pay or receive, depending on whether they are short or long. In GB, by contrast, there is a dual cash out price: market participants that are short may pay a different (and typically higher) balancing price than the price received by those market participants that are long. Since this topic is discussed explicitly later, the following section focuses on the first two aspects.

2.4 Pros and cons of market versus mechanism

What are the pros and cons of these philosophies or sets of arrangements? Is a balancing market as in Texas better or worse than a balancing mechanism as in the UK? There are arguments on both sides.

Consider first the practice of allowing a broader scope and discretion to the System Operator to buy ahead. This can yield lower costs of acquiring balancing services if it reduces risk for the generator offering the service, or if the System Operator has a more informed view than individual market participants with respect to the likely future situation in the balancing mechanism. There is a belief, and some evidence, that incentives have reduced the System Operator’s costs in GB, and thereby reduced overall prices to customers. Ofgem has shown a continued preference for incentive schemes on the System Operator (on acceptable terms).¹¹

But such an active role for the System Operator can also present problems for other market participants, particularly if relevant traded markets are illiquid. The System Operator’s “large portfolio of balancing service contracts ... has the effect of taking significant quantities of plant off-market”.¹² A less constrained System Operator may be able to trade on a more favoured basis, entering contracts with less concern about the risk of making wrong judgements, and with other parties keener to contract with it because of the lower risk of non-payment. This may be the more serious if the actions of the System

¹¹ For the latest proposal, see *National Grid Electricity System Transmission Operator Incentives from 1 April 2007*, Ref 35/07, Ofgem, 27 February 2007.

¹² Cornwall, *Cash-out Review 2007*, section 2.6

Operator are opaque, and if the cost of each purchase outside the balancing mechanism is not clearly identified and justified.

Similarly, pay-as-bid was assumed to be cheaper for buyers and hence for customers. Paying all bidders the price of the highest-price marginal unit would mean paying intra-marginal generators an unnecessarily high price. If the System Operator could act as a discriminating monopsonist this would reduce the total cost of acquiring the necessary units.

Others challenged this.¹³ It has been argued that

- under pay-as-bid the market participants would no longer have an incentive to bid their avoidable costs, but instead would base their bids on conjectures about the market price,
- this would largely remove the hoped-for savings from pay-as-bid,
- the outcome depends on the precise assumptions made, but in theory and in practice there is generally not a significant difference between the expected prices under the two approaches,
- pay-as-bid would introduce additional risks and costs associated with forecasting
- this would lead to inefficiencies since higher cost plants would sometimes bid lower than high cost plants and be chosen to run, and
- this would hinder rather than facilitate competition because the risks and costs would be relatively higher for smaller firms than for larger ones.

Some of the concerns refer primarily to pay-as-bid as a basis for pricing in a Pool, or might carry more weight in a context where the majority of output were dependent on the bidding system. This is not how pay-as-bid was envisaged or has been implemented in GB, where the main change was to bilateral contracts, and the balancing mechanism accounts for only a small percentage of total output. However, there are additional concerns as to whether pay-as-bid yields a price that is sufficiently similar to a market price to constitute a benchmark for a liquid contract market, and whether it gives market participants adequate price signals about marginal cost as opposed to average cost. These points are discussed further below.

2.5 Conclusions on balancing market versus balancing mechanism

These advantages and disadvantages of markets and mechanisms have to be balanced. There are potential advantages associated with giving the System Operator stronger incentives and wider discretion, in terms of lower expected costs of balancing. But there are potential disadvantages in terms of undermining the bilateral markets and the market participants, including those participants that trade in them as well as those that generate

¹³ E.g. Alfred E Kahn et al, *Pricing in the California Power Exchange Electricity Market: Should California Switch from Uniform Pricing to Pay-as-bid Pricing?* Blue Ribbon Panel Report, California Power Exchange, 2001. Alfred E Kahn et al, *Uniform Pricing or Pay-as-Bid Pricing: a Dilemma for California and Beyond*, *Electricity Journal*, July 2001, pp. 70-79. Peter Cramton and Steven Soft, *Why We Need to Stick with Uniform-Price Auctions in Electricity Markets*, *Electricity Journal*, Jan/Feb 2007, pp. 26-37.

and supply. The balance of advantages and disadvantages will depend upon the particular circumstances, and have surely changed over time. Four points may be made here.

First, there is a significant difference between the GB and other markets including Texas, in that the System Operator is a privately owned company in GB but not in Texas and elsewhere. Insofar as a private operator is more responsive to financial incentives, it can be made more responsive to the interests of other market participants by a suitable incentive mechanism. There are thus potentially greater benefits for other market participants from incentivising the System Operator to be more efficient in discharging its specified duties and providing agreed services – and, as discussed later, in devising new and potentially more attractive services for those market participants. On this basis, there may be a stronger case for a broader interpretation of the System Operator's role in GB than elsewhere. At least, this approach should not be lightly abandoned.

Second, it is nonetheless necessary to recognise that a broad role for the System Operator raises important issues of design, monitoring and evaluation. Can the System Operator's conduct be sufficiently well specified and evaluated that tangible benefits can be actually identified? And can the System Operator's conduct be sufficiently well constrained and identified that market participants are not unduly disadvantaged? There seems to be some concern that the actions of the System Operator are not as transparent or as speedily declared in GB as they might be, or as they are in some other electricity markets. It is noticeable that the design of the incentive arrangements is primarily for Ofgem to take forward, rather than for market participants generally (though there are opportunities to comment on consultation papers). It is difficult to evaluate whether the incentive arrangements have been set at the right level, and have produced as much value as they might have done, and have not distorted the development or prices in the markets for other electricity products. The relationship between the actions and interactions of a single party acting as System Operator and Transmission Operator may also be an issue.

Nigel Cornwall's paper provides some useful suggestions on accountability. For example, it might be helpful to require further, earlier and more specific identification or tagging of trades undertaken by the System Operator for each of its functions, to the extent that this is feasible with present computer systems.

Further steps could be considered to address accountability issues. Independent Market Monitors in the US and in Latin America are beginning to monitor and evaluate the impacts of System Operator actions on market efficiency. This experience would seem worth reviewing and considering for GB. It may be possible for the governance arrangements to encourage a more active role for market participants as well as, or instead of, Ofgem, as discussed later in this paper. A more thorough process of ex post analysis by representatives of the market participants would seem useful, with audit reports made available to market participants. Some would argue that, because the System Operator, Ofgem and market participants are all interested parties, completely independent auditors should be appointed to review and evaluate the actions of the System Operator at frequent intervals. They could also suggest or comment on possible improvements to rules, procedures and software.

Third, to the extent that some of the concerns that influenced the initial NETA cash out arrangements no longer apply with such force, it seems worth reconsidering the balance of advantages between a market and a mechanism. For example, to the extent that the System Operator discretion and the pay-as-bid features reflected concerns about generator market power or the capability of the System Operator, these are no longer such concerns. They should not preclude the use of more conventional arrangements that would be preferable on other grounds. Arrangements whereby System Operators have strictly limited powers of discretion and calculate market-clearing balancing prices rather than pay-as-bid seem to work well in various markets in the US and in Latin America.

Fourth, although a balancing mechanism may be seen as responding more flexibly to the needs of market participants, without restricting the System Operator to providing a market, it is quite possible that the provision of a market is a particular kind of service that many market participants would find helpful. For example, a concern when NETA was introduced was to facilitate a more liquid short-term market. In the event, liquidity has improved in some respects but deteriorated in others. It has been said that the absence of a balancing market as exists elsewhere has been one of the problematic factors in GB. If this is the case, then there is a stronger argument for a more conventional balancing market rather than for a more discretionary balancing mechanism.

To summarise, an active and discretionary role for the System Operator was considered appropriate when NETA was introduced. It seems to have produced benefits in GB (or at least in the previous England and Wales system), and should not lightly be discarded. There are potential advantages in incentivising the System Operator continually to seek out and provide balancing and other services that market participants find attractive and efficient.

However, there are also potential disadvantages with a balancing mechanism that involves greater discretion for the System Operator and also with a pay-as-bid approach. It is difficult to judge the continued effectiveness of this mechanism, and the case for it no longer seems as strong as it might have been initially. There would be advantage in requiring greater accountability and auditing of the System Operator's actions. Independent market monitors are doing this in electricity markets elsewhere. In addition, there would now seem to be greater advantage to market participants, and hence ultimately to customers too, in developing a more conventional balancing market. This would have less discretion for the System Operator and would not have pay-as-bid as the basis for setting cash out prices. It would give greater weight to the role of a balancing market as enabling market participants to trade with each other.

3. Single versus dual cash out price

It might have been expected that the new electricity trading arrangements (NETA) would be set up with a single cash out price – that is, with System Buy Price (SBP) equal to System Sell Price (SSP). Yet in the event a dual cash out approach was adopted. SBP and SSP would in principle be calculated separately and would normally be different, although in some circumstances they could be the same.¹⁴

Most other electricity markets have not adopted dual cash out prices, or at least not on the same basis as in GB.¹⁵ What were the arguments for dual cash out, and do they still remain valid?

3.1 Market power and assisting the System Operator?

I have noted above the concerns about generator market power in the Pool, and the associated ability to influence Pool price. This concern may have discouraged the idea of a single cash out price under NETA, because that price could conceivably come to replace Pool price as a vehicle for manipulating the market.

I have also noted concerns about the System Operator's ability to balance the system under the new arrangements. This may have favoured a dual cash out price in order artificially to increase the incentives on market participants to balance. It is sometimes said that the difference between the two prices is effectively a tax on out-of-balance participants.¹⁶

I have suggested that these concerns could have been more valid when generator market power was a bigger problem, and in the early and uncertain days of a new trading system. But they are less serious concerns now, and should no longer drive the design of trading arrangements today and for the future.

It is in any case unclear whether a dual cash out price does make it easier for the System Operator to balance the system. A high SBP when the system is long is indeed likely to discourage market participants from being short – but a low SSP is likely to discourage market participants from increasing the length of their position. In other words, a low SSP tends to undermine the beneficial incentive effect of a high SBP, rather than reinforce it. Any incentive to system balance may thus derive primarily from the size of SBP rather than from the difference between SBP and SSP.

¹⁴ The initial Proposals envisaged a single balancing price in each period. *Review of Electricity Trading Arrangements: Proposals*, Offer, July 1998, (e.g. para 5.24). The 'two price cash-out regime for imbalances' appeared in later more detailed proposals. *The new electricity trading arrangements, Volume 1*, Ofgem, July 1999, section 7.1.

¹⁵ Of the ten markets mentioned by Cornwall (*Cash-out Review 2007*, section 2.1), only Nordpool (excluding Norway) has two cash out prices conceptually similar to GB. France and Netherlands each has a single price adjusted by premia and discounts.

¹⁶ Cornwall, *Cash-out Review 2007*, section 1.2. He also remarks "The SSP-SBP methodology is premised on a simple concept – the desire to encourage parties to balance their positions *as an end in itself*."

3.2 Different costs?

The objective of cash out arrangements that does remain valid is that of facilitating the development of a competitive and efficient electricity market. To that end, cash out prices should reflect the costs incurred by the System Operator in balancing the system. Market participants would then decide how far to balance their own positions and how far to use the System Operator via the cash out mechanism, depending on the relative costs of each approach.

On this basis, there could still be a justification for dual cash out prices if the System Operator's costs of dealing with participants that are long in any period were systematically and significantly different from the costs of dealing with participants that are short in that period. Or, put another way, the use of dual cash out prices presumes that the costs incurred by the System Operator in balancing the system depend on the specific positions of each of the parties and not on the net position of the system as a whole.

In principle this could be true. It might be possible to instance specific examples – either hypothetical or practical - where this is the case. But I have not seen any convincing theoretical argument nor any empirical demonstration to show that it is a significant and persistent feature of the balancing process as it actually applies in GB. This is not surprising: during real-time operation a System Operator balances the system as a whole. It cannot normally identify imbalances on the part of individual market participants, which are only revealed in the subsequent settlement process.

Other aspects of the balancing mechanism calculations seem inconsistent with the proposition that individual positions rather than the overall net position are most relevant. For example, since mid-2003 trades in opposite directions within each half hour period are netted off and only the net trades determine cash out prices. If the System Operator's costs are not determined by the net imbalance position, then the whole basis of setting cash out prices based on the Net Imbalance Volume (NIV) is called into question.¹⁷

3.3 A moral justification?

A variety of other justifications have been adduced for maintaining the dual cash out regime. For example, it has been suggested that “parties that are long when the market is short ... are not in any meaningful sense contributing to balancing the system (except

¹⁷ Ofgem's P 74 Decision Letter says “While it is difficult to value the actual cost imposed by the Party being out of balance, to assume that the cost is zero by adopting a single cashout price would be even more arbitrary.” I do not understand this sentence. The cost imposed by a party being out of balance (short) is surely given by the SBP in either case, not by the difference if any between SBP and SSP. A single cash out price, calculated properly, would fully reflect the actual cost imposed by a party being out of balance. The Decision Letter continues “Consequently, it is appropriate that participants who are spilling electricity should receive a lower price for their electricity than if they had been fully contracted since they may be imposing costs on the system.” The argument above is that it has not been convincingly demonstrated that this is the case. In the absence of evidence to the contrary, the presumption underlying the calculation of the present cash out price in the main direction is that participants who are spilling electricity provide a benefit to the system equal at the margin to the cost imposed by those who are short.

inadvertently)".¹⁸ This seems to invoke a moral justification for paying some participants for some services but not others. This seems questionable when there are invariably imbalance elements on the system in both directions. Is the suggestion that the System Operator should judge and reward those market actions that are meaningful and deliberate, while penalising those that are meaningless and inadvertent? Should it reward more highly a generator for bidding to operate long than a generator that decides to operate long, if both make the same contribution to dealing with net imbalance volume in the system as a whole?

Such moral distinctions seem difficult and arbitrary to make. They also have economic costs. Insofar as a difference between SBP and SSP in any period is a source of inefficiency and not conducive to competitive markets, then it is a measure of the cost of the moral stance apparently being taken.

3.4 Short term traded market prices?

The initial definitions of SBP and SSP were based on the costs incurred by the System Operator, which was the stated philosophy of the cash out prices. In mid-2003 modification P78 set the reverse cash out price equal to a market price based on short-term energy trades made in the forward market.¹⁹ If the reverse price has been 'deliberately delinked' from the System Operator's costs is difficult to see how this is consistent with the stated philosophy of setting cash out prices to reflect the System Operator's costs.

It is possible to think of other arguments or philosophies for setting cash out prices. Reverse cash out prices based on forward market prices may be more attractive to those participants that are usually long, simply because they are higher than those based on the previous method of calculation. They may be more stable and predictable. They may usefully reduce the obligation on the System Operator to record, allocate and aggregate relevant costs of dealing with imbalances. They might be implied by a philosophy that says 'regardless of what the System Operator's costs actually are, it would be desirable for market participants to be able to cash out imbalances at short term energy market prices that are independent of the System Operator's actions and judgements'.

Although arguments could be advanced along these lines, they would not embody the philosophy that is claimed to apply in the design of GB trading arrangements. Forward market prices are not the costs incurred by the System Operator. These are not the costs that are supposed to be reflected to market participants, on the basis of which they make their decisions how far to balance, and on the basis of which the system is supposed to lead to an efficient extent of actions by the System Operator.

3.5 Does it matter whether there is a dual or single cash out pricing system?

¹⁸ *Electricity and gas cash out review, A Consultation Document*, Ofgem, May 2004, para 2.22, p. 15.

¹⁹ 'Reverse' refers to individual imbalances in the opposite direction to that of the system as a whole, which is referred to as the 'main' direction.

Some might argue that dual versus single cash out – and perhaps cash out generally - is no longer a significant issue. It might be said that the average gap between system buy and sell prices has reduced over time, and is now considerably less than it was initially. Or it might be argued that cash out trades and revenues are a relatively small part of the whole picture. Cash out volumes are less than 5 per cent of total electricity supplied. Major decisions about generation investment and supply are more oriented to the 95 per cent of volume and associated costs than to this 5 per cent. Moreover, under the pay-as-bid arrangement generators that are chosen to run already receive their bid price, so the present system does not necessarily represent a disincentive to them.

It is true that the tail should not wag the dog. The unrestricted nature of the GB generation and supply markets means that the dog is relatively healthy. The major concerns as regards investment in GB perhaps relate to other factors such as uncertainties about government policy on (e.g.) climate change, renewable energy and nuclear energy. On this basis, the way that cash out services are priced is a second-order determinant of major investments in capacity.

Nevertheless, there are several reasons why the dual versus single cash out issue does matter.

- First, the various electricity markets and mechanisms are inter-related. Prices in the balancing mechanism are likely to influence spot market prices which in turn will influence longer term contract prices. This may not be so apparent in GB, where there is a more complex dual price rather than a single cash out price. But it may be clearer in other markets. In Texas, for example, the general perception is that higher than competitive BES prices in summer 2005 significantly impacted on bilateral prices in the ERCOT wholesale market.
- Second, although steps have been taken to reduce the gap between SBP and SSP, this does not guarantee that the gap will remain small. It has actually increased recently and is quite significant. According to a recent calculation, SBP averaged 17.8% above spot price in 2006 while SSP averaged 18.6% below it, an average difference of about 35% of spot price.²⁰ Moreover, the average gap masks what is happening in different situations, and to different market participants. And while cash out trades are indeed a relatively small part of the whole picture, they can have a significant impact on market decisions. So that is not an argument for refraining from putting balancing arrangements on a more efficient and productive basis.
- Third, the significant gap between system buy and sell price seems likely to have distorted decisions on how far each participant decides to balance its own position rather than use the facilities of the System Operator. Not surprisingly, many market participants seem to have taken the view that being short is to be avoided at almost all costs. This is unlikely to be efficient. The dual price makes unduly expensive an option that may in fact be lower cost, especially for smaller

²⁰ Cornwall, *Cash-out Review 2007*, section 1.2.

participants, than the use of traded power exchanges. It is also conceivable that it may have encouraged the distinctly more vertically integrated structure of the industry in the UK compared to some other markets.

- Fourth, the gap between SBP and SSP price is unlikely to assist the System Operator as fully as some have claimed. This is most importantly the case at times when the system as a whole is short, as noted above. SBP may be particularly high, but SSP may fall far below this. Those market participants that are short are effectively encouraged to moderate their short positions. But those market participants that are long are not effectively encouraged to lengthen their long positions. Both types of participants can make an effective contribution to balancing the system, but the dual cash out mechanism encourages only one set of market participants to do so. In fact, insofar as market participants are unsure whether they are likely to be short or long, the calculation of SSP undermines rather than reinforces the incentive provided by SBP.
- Fifth, the effect of the dual cash out arrangement, coupled with the Residual Cashflow Reallocation Cashflow (RCRC) provisions is to divorce the costs and revenues of the balancing mechanism. It is no longer the case that total revenues equal total costs in each period. In practice there has been a surplus (tellingly known as the beer fund). As argued later, this is likely to have an adverse effect on the accountability of the System Operator.
- Sixth, this mechanism in turn systematically tends to favour the larger and more vertically integrated market participants over the smaller non-integrated ones. This is because the smaller participants use the cash out mechanism proportionately more than larger ones, perhaps partly because they are not so able to offset volumes in each direction. But the beer fund revenues are distributed in proportion to size of participant. Smaller participants therefore pay in a greater proportion of the net revenues of the beer fund than they take out.²¹ There is more on the disadvantages of the RCRC approach below.
- Seventh, the absence of a single cash out price seems to have made it more difficult for market makers and traders to establish a product that can be widely traded. In Texas, for example, trades have been made based on MCPE. In GB, by contrast, trades are not based on cash out prices. More generally there have been concerns in the GB market about the lack of a suitable liquid reference price. It is widely said that a single cash out price would constitute such a reference price whereas the present dual price does not. Whether this would make the difference in creating a more liquid market is more difficult to judge, and I understand that other possibilities are under consideration. But it would be very unfortunate if

²¹ This at least seems to have been the position in late 2005, when I made some unpublished calculations. See also my Report on *Smaller Suppliers in the UK Domestic Electricity Market*. Some of the smaller suppliers have since left the market, but the situation reportedly still obtains (Cornwall, *Cash-out review 2007*, sections 2.1, 3.4).

dual cash out reduced or removed the prospects of developing a more liquid short term trading market.

- Eighth, the present arrangement is that reverse cash out price is delinked from the System Operator's costs and related to a traded market price. This runs the risk of distorting the traded market by introducing incentives to influence that market price in order to influence cash out prices.

3.6 Conclusions on dual versus single cash out

There might have been defensible reasons for dual cash out in the early days of NETA, with a view to curbing generator market power and ensuring the System Operator's ability to balance the system. To the extent that they were valid, those reasons no longer apply with such force today. The result is that there appears to be no logical, consistent and defensible explanation for continuing to maintain dual cash out. This undermines the credibility of the case for the UK's general approach to electricity markets - that is, the case for a bilateral trading system with a balancing mechanism geared to encouraging efficiency and facilitating competition in the electricity market.

The dual cash out system has other disadvantages. It seems likely to distort the short-term decisions of market participants on the extent of cover in each period, and the long-term decisions on the extent of balancing and vertical integration. It may systematically favour larger integrated players compared to smaller ones. It divorces cash out revenues from costs; this is not conducive to the accountability of the System Operator. And a dual price system is not conducive to a liquid short term trading product.

Conversely, a single cash out regime would be more understandable and defensible. It would be less distorting and would improve incentives on market participants. It would provide a more economic option for smaller market participants and new entrants. It would be more consistent with improved arrangements for accountability of the System Operator. And it would provide a more attractive basis for a short-term trading product and a more liquid short-term market.

Some might be concerned that a single cash out price arrangement would increase the risk of the System Operator not being able to balance the system. However, there is no reason why this should be the case, especially given the flexibility accorded to the System Operator under present arrangements. But experience elsewhere suggests there would not be serious problems even without the present flexibility. There is no reason to expect that market participants would persistently over-react. A single cash out price has not proved a problem in other market jurisdictions.

It would therefore seem sensible to convert the dual cash out arrangement to a single cash out arrangement. The System Operator would set a single cash out price in each period. This would be a relatively high or low price depending upon whether the system as a whole was short or long. It would be paid by those that are short and received by those that are long. This would be more conducive to competition and economic efficiency than

the present approach. It would also bring GB arrangements more closely into line with best international practice in competitive electricity markets.

4. Cash out prices based on average cost versus marginal cost

From the beginning of NETA until very recently, GB cash out prices have been based on average cost. In simple terms, the imbalance price in each half hour was calculated by ranking in increasing order of cost the purchases made by the System Operator to balance the system. After various adjustments the cash out price was set equal to the average cost of these actions. In contrast, in most other electricity markets, imbalance prices have been based on marginal cost – the cost of the last (most expensive) unit purchased by the System Operator.

Over time, there have been arguments for setting cash out prices equal to marginal cost – say, the cost of the System Operator’s most expensive purchase in that half hour (again, after making various adjustments). There has recently been a move in this direction. Modification P194 set the cash out price equal to the average cost of the most expensive 100 MWh of purchases (so-called ‘chunky marginal’). Then in October 2006, just before P194 was actually put into practice, modification P205 provided that cash out price should be set equal to the average cost of the most expensive 500 MWh of purchases.

I previously suggested a means whereby cash out prices could better reflect marginal cost in a different way, consistent with just covering total cost. However, the case for this variation assumed a continuation of dual cash out prices and pay-as-bid. Suppose both of those aspects of the cash out arrangements are modified, as suggested above. And suppose that the rules of cost allocation and tagging are modified, as discussed below. Then the case for marginal cost pricing versus average cost pricing, and for the scheme I suggested previously, all appear in a different light.

It is convenient to begin by summarising the arguments hitherto made in GB for and against setting cash out price equal to marginal cost or average cost. A significant concern has attached to unpredictable or ‘polluted’ marginal prices. There is an additional argument in the economic literature, about the importance for accountability of just covering total costs. I briefly review my suggested scheme for combining marginal cost pricing with breaking even under the present arrangements. Then I explore whether there needs to be a divergence between total revenue and total costs with marginal cost pricing. This leads back to the question of predictable or ‘polluted’ bids, which in turn leads on to the discussion of cost allocation and tagging in the next section.

4.1 Average and marginal cost pricing

There is a long-standing debate in economics about the merits of average cost pricing versus marginal cost pricing. A widely held view is that marginal cost pricing is economically ‘correct’ and will lead to the efficient allocation of resources. According to this view, approaches other than marginal cost pricing – for example, average cost pricing – are correspondingly ‘incorrect’ and inefficient.

Marginal cost pricing has the obvious advantage that it reflects cost at the margin. It therefore gives a more accurate picture to a supplier or generator as to the costs or benefits of trying to adjust the extent of its imbalance. When capacity is tight and the cost of balancing the system rises sharply with the extent of imbalance, a marginal price would be significantly higher than an average price. This would presumably give a greater incentive to market participants to avoid being out of balance. It is also argued that failure to price on a marginal basis gives inadequate incentive to generators to install, maintain and offer the services of relevant generation capacity, such as peaking and rapid response capacity. From this perspective the failure to set cash out prices equal to marginal cost was or still is a weakness in the present system, at least until modified by P194 and P 205. It is therefore understandable that the System Operator should argue for cash out price to be set on a marginal basis. The System Operator was indeed the proponent of modification P194.

The original Proposals for NETA noted that arguments and support for marginal pricing versus pay-as-bid were about evenly balanced, and discussed the pros and cons of each approach. In coming down in favour of pay-as-bid it made an interesting point on the efficiency criterion.

The balancing market will be open for several hours, including real time operation. During this period conditions on the system will be continuously changing. Trades may be accepted at particular times at prices that are quite different from the average price of accepted trades over the period as a whole. Consequently, there is no obvious definition for the marginal or market clearing price throughout the period. To pay a uniform accepted price on all increments of generation and decrements of demand, which would presumably have to be the highest price accepted from any one of them, would not obviously be more efficient and could be expensive. For similar reasons, it is not obvious that imbalance prices set on a marginal basis would be more efficient than those set equal to average cost.²²

The weight to be attached to this point will of course depend on the time of gate closure (which has since been reduced from three and a half hours to one hour in GB) and on the duration of the trading and settlement period (which is now fifteen or even five minutes in some overseas markets compared to 30 minutes in GB).

Later resistance to marginal cost pricing of cash out prices in GB has focused on certain practical disadvantages (also recognised earlier). It may lead to erratic prices, especially where the price order 'stack' can be distorted or 'polluted' by system actions taken by the System Operator. This may be a more serious problem where the 'stack' contains a large

²² *Review of Electricity Trading Arrangements: Proposals*, Offer, July 1998, para 4.49. The Proposals envisaged that "A pay-as-bid process would seem to have advantages in encouraging competition" (para 3.12) because generators would actively have to propose a price that would be sufficiently competitive, rather than passively receive the going market price. It was explicitly recognised (paras 3.13 – 3.16) that this approach would not be appropriate in a day-ahead auction like the Pool. Incidentally, the Proposals refer consistently to a 'balancing market'.

number of actions related to transmission constraints and intra-period voltage fluctuations, than under arrangements in other electricity markets where the stack is largely limited to balancing actions. The marginal price may also be susceptible to manipulation where there are relatively few offerors and bidders.

Average cost pricing has the disadvantage that it does not reflect cost at the margin. It may be considerably less than marginal cost when capacity is particularly short. On the other hand it may be more stable and less subject to manipulation.

Prices determined by modifications P194 and P205 are intermediate between average and marginal cost prices. They therefore share the advantages and disadvantages of each. P194 prices (based on the highest price 100 MWh) are closer to a marginal price and P205 prices (based on the highest price 500 MWh) are closer to an average price.

Ofgem had to weighing the pros and cons of each approach. It concluded that both modifications gave improved signals to balance, relative to the previous approach based on average cost. It considered that the benefits of less potential price distortions via the deeper stack of P205 compared to P194 outweighed the potential detriment of any reduction in the signal to balance.²³

4.2 Marginal cost pricing and covering costs

There is another view in the economic literature, that marginal cost is important but only part of the whole picture. All methods of pricing have both advantages and disadvantages. The challenge is to find the method that is the most advantageous or least disadvantageous, taking all considerations into account.²⁴

Marginal cost pricing does not generally lead to total revenues that just cover total costs. In the case of declining cost industries or services, for which marginal cost pricing was originally advocated, the consequence would be that the enterprise would make losses. This could reduce accountability and efficiency within the enterprise. Funding the deficits out of taxation would introduce additional distortions.

With increasing cost industries or services, the enterprise would recover more revenue than total costs. This too could reduce efficiency and accountability. The distribution of the resulting surplus may also be problematic and distorting.

Cash out services are presumably in the increasing cost category. In principle the System Operator ranks and selects the bids and offers in terms of increasing cost. With pay-as-bid, a cash out price based on the last (highest) bid accepted would be higher than one

²³ For a more extensive account see Cornwall, *Cash-out review 2007*, section 2.2.

²⁴ See especially R H Coase, The marginal cost controversy, *Economica*, n.s. 13, August 1946, reprinted in R H Coase, *The Firm, the Market and the Law*, University of Chicago Press, 1988, pp. 75-93. R H Coase, The theory of public utility pricing and its application, *The Bell Journal of Economics and Management Science*, 1 (1), Spring 1970, pp. 113-128.

based on the average bid accepted. If cash out prices were based on marginal cost, this would lead to a surplus of revenue over total cost.

In practice, the situation is more complicated because of the alternative ways of providing balancing services, the various rules for allocating costs and tagging (removing) certain bids and offers, the dual price basis of cash out, and the different means of recovering revenues. The incentive schemes on the System Operator may also have added complexity. These factors have led to a situation where the relationships between the prices and costs of individual services, and between total revenues and total costs, have become very blurred.

A major flaw seems to be the disjunction between the revenues for the provision of balancing services and the costs incurred. In each period there are rules for determining prices, but there is no requirement that the prices lead to revenues equal to the costs. Instead, the difference is simply distributed to or recovered from market participants according to the RCRC mechanism.²⁵

At various times the discrepancy between total cash out costs and total cash out revenues has been considerable. It has increased in the past year with the higher costs of energy. The discrepancy is funded disproportionately by smaller market participants, whose imbalance volumes usually constitute a larger proportion of their total electricity purchases. Not only is the present basis of cash out prices not cost-reflective, the mechanism for rebating the surplus further distorts cost-reflectivity. And if the System Operator is assured of covering its costs (subject to the incentive mechanisms) from one pot or the other, that would seem likely to reduce its concern about the efficient provision and proper pricing of the components of each service.

4.3 Accountability

Present arrangements are thus not conducive to holding the System Operator fully accountable for its actions and inactions in providing imbalance services. It is important that market participants and others can make a judgement about the value of each of the services provided and the efficiency with which it is done.

What conditions are most conducive to such accountability? It would seem helpful to require that the services are well specified, and that each service is priced such that the associated revenues recover the costs of providing it. The System Operator may well be providing a range of services – including, for example, the provision of transmission, the management of transmission constraints and the provision of energy and system balancing services. It is important to be able to identify which actions are taken as part of the provision of each service, what are the costs of each action and what are the associated revenues. Such ‘direct assignment’ of costs is regarded as important in electricity markets elsewhere.

²⁵ “In fact the actual cost of imbalances is recovered not through the imbalance prices at all but through the mechanism of BSUoS, which smears the actual costs of balancing the system across all grid users.” Cornwall, *Cash-out Review 2007*, section 2.3, also 2.4.

It will then be easier to decide whether the value of each service justifies its provision, whether it should be modified or repriced, whether it is possible to provide the same service with greater efficiency, and so on. Not least, it is important that market participants should be able to judge whether it is better to use a 'market' to provide balancing services, as in other jurisdictions, or a 'mechanism' embodying the kind of discretion for the System Operator that has characterised the GB approach to date.

Of course, it may happen that in practice costs and revenues turn out to be different than expected, and over- or under-recover in particular respects. This would therefore need to be explained and justified.

There is also the question of incentive schemes on the System Operator. It could be allowed, as at present, to earn revenues in excess of cost, and keep a proportion of that excess, if it managed to reduce the costs of provision as a result of superior efficiency. (Under such schemes the System Operator would presumably also run a corresponding risk of not covering its costs if it failed to meet the efficiency targets.) However, the decision to introduce such a scheme should take into account the implications for transparency and accountability, as well as for efficient purchasing.

4.4 Quantity discounts and quantity premium bands

There is thus an advantage in requiring that total revenues for cash out services should cover total costs. There is also an advantage in cash out prices reflecting costs at the margin, which in practice in GB has led to revenues considerably in excess of costs. How are these two principles to be reconciled?

On the basis that marginal cost pricing of cash out would continue to lead to a surplus of revenues in the GB context, I previously explored an alternative way of resolving this problem.²⁶ Commercial businesses deal with an analogous problem of decreasing costs, by means of quantity discounts. The base price reflects the cost of doing business with smaller quantities, but discounts are available to reflect the lower costs of larger quantities. This enables the business to charge a lower price at the margin that is closer to marginal cost and thereby enables the business to achieve what economists would call an efficient level of output, while still covering total costs. This suggested that it would be possible to deal with the problem of increasing costs by means of quantity premia.

Modifications P201 and P202 introduced the concepts of tolerance bands, which in effect embodied different cash out prices for different levels of imbalance. It is possible to use this idea as a basis of setting cash out prices. The cash out price for the first 20 MWh (say) of imbalance per market participant could reflect the System Operator's costs of meeting imbalance requirements up to that level. A higher price – in this case a quantity

²⁶ Stephen Littlechild, Imbalance prices, tolerance bands and quantity premium bands, *Energy Spectrum*, Cornwall Energy Associates, 4 September 2006, also available at <http://www.electricitypolicy.org.uk/pubs/misc.html>.

premium – could be charged for amounts above that, reflecting the costs of meeting imbalances above 20 MWh per market participant.

On this basis, 1) all prices charged would reflect the costs incurred by the System Operator, 2) the total revenues would (just) recover total costs, and at the same time 3) the cash out prices would more accurately reflect costs at the margin than a single average cost price would. Detailed rules would be required, but the principle is clear. It might be argued that such cash out prices would be more complex than present prices. But the present basis of price setting, with its dual prices and tagging and other procedures, is far from simple.

Such cash out prices would seem to constitute a better and more cost-reflective service for smaller market participants, who would be less exposed to the fluctuations in imbalance volumes and cost that are primarily associated with the fluctuations in imbalance volumes of the larger players. This would therefore be conducive to more competition. The larger players would gain from the improved incentive to balance – and hence the improved security in the system - associated with the higher cash out prices for higher imbalance volumes. At the same time there would be no discrimination against or in favour of any size of market participant. Quantity premium charges would seem to reflect costs more accurately than single or dual cash out prices.

4.5 Is marginal cost pricing necessarily inconsistent with just covering costs?

This quantity premium proposal presumed that marginal cost pricing of cash out would lead to revenues exceeding costs in the GB context. But is that necessarily the case? And why does this not seem to be an issue in other electricity markets?

Two aspects of the GB approach seem to be relevant here. First is the dual cash out policy. If shortages are priced at the marginal cost of providing additional power and surpluses are priced at the marginal cost of absorbing power, where these are assumed in general to be different, then there is no reason why total cash out revenues should generally equal total cash out costs. There is the additional reason that, since modification P78, reverse cash out price has explicitly been set equal to a market price that is not claimed to equal the System Operator's costs.

The second relevant aspect of the GB approach is the pay-as-bid policy. If bids and offers into the balancing mechanism are paid at bid price, but those who take balancing services pay the price of the highest unit accepted, then total revenues will not generally equal total costs.

But suppose there were a single cash out price instead of a dual cash out price, as suggested above. And suppose that instead of pay-as-bid, a uniform price was paid for all bids and offers in each period. In that case a marginal cost price would indeed lead (at least in principle) to total cash out revenues just covering the total costs of balancing the system, in each period. This is because the price charged by the System Operator for each

imbalance unit, whether long or short, would equal the cost that the System Operator paid for it.

This puts the debate about average versus marginal cost pricing of cash out into a different light. Suppose that changes in cash out arrangements mean that marginal cost pricing no longer implies a divergence between total revenues and total costs. Then an important part of the case for marginal pricing – and indeed of the case for moving from pay-as-bid to paying a marginal price - turns upon the considerations related to predictability and susceptibility to manipulation. This in turn leads on to questions of cost allocation and tagging of System Operator trades.

4.6 Conclusions on average versus marginal cost cash out prices

GB cash out prices have traditionally been set equal to the average cost of providing balancing services. There have been arguments for a move to marginal cost pricing, particularly to ensure sufficient cover at times of system peak. The main concern has been vulnerability to unpredictable and manipulated prices. Recently it has been decided to adopt ‘chunky marginal’ cost pricing.

I have suggested that the ability to hold the System Operator accountable for the services it provides, and its prices and efficiency, is an important consideration. This would be assisted by a framework in which the System Operator’s total revenues should equal its total costs, including for each individual service separately. If marginal cost pricing would otherwise preclude this, there would be merit in a system of cash out prices with quantity premia, to enable cash out prices to better reflect costs at the margin at times of particular system stress, while still just recovering total cash out costs.

A disjunction between cash out revenues and costs is implied by the present policy of dual cash out prices and by the pay-as-bid rule. If these were modified, marginal cost pricing of cash out would not be inconsistent with breaking even. The case for a move to marginal cost pricing, both in setting SBP and SSP and in paying the providers of bids and offers, therefore turns on the possible impact on considerations such as predictability and vulnerability to manipulation. These in turn are influenced by policy on cost allocation and tagging.

5. Cost allocation and tagging

Since the implementation of NETA there have been concerns about the precise definition and calculation of cash out prices. Numerous revisions have been proposed, and some have been adopted. This section looks at the approach to cost allocation and tagging.

5.1 Cost allocation

Since its inception, the GB system has generally been quite long. At present it is long about three quarters of the time.²⁷ This is not necessarily problematic, though it may not be entirely economic. A more pressing concern recently is that the system is not generally long at times of peak demand. At such times Net Imbalance Volume is around zero. This suggests that the incentives on market participants to balance their positions are less strong at such times. Some would say they are insufficient.

Various possible reasons are put forward for this. One is that SBP is set equal to average cost rather than marginal cost, or at some point in between. The difference between average and marginal cost is more marked at times of peak output. If marginal cost is taken as the appropriate benchmark, then balancing services are particularly underpriced at times of peak. Market participants are artificially encouraged to rely on the System Operator at such times rather than make their own arrangements. The case for moving from average to marginal cost pricing has been discussed above.

Another possible explanation relates to the present cost allocation rules. It is said that cash out prices at peak times do not properly reflect the costs of providing balancing energy at peak times. Two examples of this may be given.

The first example is where the System Operator wishes to buy energy for one or two consecutive peak half-hour periods. Because of plant dynamics, the generator may need to operate (including starting up and shutting down) over more than two consecutive periods. At present the System Operator 'honours' these plant dynamics by contracting for several consecutive periods. But only the costs incurred in the one or two peak periods appear in the cash out prices for those periods. The costs of output (or potential output) in the other periods appear in the cash out prices for off-peak and shoulder periods. The consequence of this rule is that cash out prices understate the full costs incurred to provide energy in peak periods, and overstate costs in other periods.

This could be remedied in at least two possible ways. One way would be to continue to purchase on the present basis, but to allocate to the peak period(s) all or a greater proportion of the total costs of contracting for this plant. The other way would be to contract and pay only for output in the peak periods, leaving the generator to cover its costs in other periods. This might imply increasing the price that it charges in peak periods or bidding into the market in the off-peak periods in the back of its peak period contract.

²⁷ Cornwall, *Cash-out Review 2007*, section 2.4.

A second example is that plant may be contracted to provide balancing services for a period of (say) several months ahead. The total costs of this are presently allocated over all the periods in which it provides services. This allocation is on the basis of a historical pattern of usage, since the actual pattern of usage of this plant is not yet determined. A concern is that the historical pattern may not be appropriate to the present usage.

However, this does not seem to be the main limitation of this approach. If the main purpose of entering this contract was to provide output over a broad set of peak periods, then all or most of the costs of the contract needs to be allocated to those peak periods. Once the contract is in place, it is economic to use it to provide output in off-peak periods, but those off-peak periods should not bear a proportionate share of the total costs. In consequence, once again the cost of providing balancing services in peak periods is understated, and the cost of providing balancing services in off-peak periods is overstated.

This distortion could presumably be remedied by revising the principles of cost allocation applicable to such contracts. With this and the previous example, it would seem appropriate to follow the general principle of peak load pricing (which is a particular case of the principle of pricing joint products). All or most of the costs of contracts that are entered into before gate closure, with a view to providing cover in peak periods, should be allocated to those peak periods. A part of the costs can be allocated to off peak periods. But no period should have to pay a higher share of the cost of a contract than the cost at the margin in the balancing mechanism in that period. And the sum of the shares on this basis should cover the total cost of the contract. Consequently, a contract should only be entered into if the resulting total cost is less than the cost of providing cover from the balancing mechanism itself.

5.2 Tagging

A different argument is that, in other respects, present procedures may overstate the costs of providing balancing services relative to the costs of providing other services. For example, at present the System Operator may take an action (for example, to remove or reduce the output of a generating plant in a particular location) primarily with a view to resolving congestion and dealing with a transmission constraint. This action may at the same time deal with an out-of-balance situation (for example, where the system was long and the System Operator would in any case need to reduce the output of a generating plant somewhere in the system). It seems that the costs of such an action are at present included in the 'stack' and contribute to the calculation of the cash out price.²⁸ This is the case even though the cost of this action may be greater than the cost of an action geared solely to dealing with the imbalance situation. In consequence, cash out price is higher than it otherwise would have been.

²⁸ Cornwall, *Cash-out Review 2007*, sections 2.2 and 2.3. Note for example that in 56 out of 63 periods in a particular data set, National Grid had taken actions to resolve system constraints that subsequently appeared in the energy stack. Note also Ofgem's repeated concerns about the imperfections of tagging and about pollution of balancing costs by actions taken for system reasons.

Where an action produces a 'joint product', the allocation of its cost among the two (or more) joint products may necessitate looking at the value of those individual products and at what it would have cost to provide them in an alternative way. It would not seem appropriate to charge the whole cost of such an action to the balancing mechanism (and hence to cash out price). The general principle enunciated above suggests that it would not be reasonable to allocate any greater amount to balancing costs than is paid at the margin for other balancing services in the balancing mechanism in the relevant periods. The same applies to the other joint products, such as the relief of transmission constraints. In total the sum of the allocated costs should cover the total cost of the action, and the action should only be undertaken if (prospectively) it does so.

How far it is possible to identify and measure the costs of (e.g.) relieving transmission constraints in different ways is obviously a consideration. Whether the cost of providing cash out services in the balancing mechanism should be calculated including or excluding the balancing volume provided by the action in question, or at some value midway between the two, is also for consideration. And how far such valuations are feasible given the software available to the System Operator is relevant. Nevertheless, there would seem a strong case for not including in the cash out price the full cost of actions taken primarily for other purposes, notably to relieve transmission constraints.

On this basis there is an argument for tagging out all such action in the calculation of cash out prices. Insofar as transmission constraints have become more significant as NETA has transitioned to BETTA to include Scotland, and are likely to be even more significant with changing technologies in future, this is a particularly important issue.

It might be argued that, if all System Operator actions that had two (or more) justifications were tagged out, then there would be more actions tagged out than left in the stack, and this could lead to a thinner basis for setting cash out prices. The implications for price setting therefore need to be examined. The exclusion of expensive constraint actions presumably means that the resulting cash out prices would generally be lower than at present, and less likely to be a source of concern to those paying them. This could facilitate a move away from pay-as-bid towards market clearing prices.

There might be concerns about the possibility and attractiveness of attempts to manipulate this price. Generator market power is not such a concern as it once was. But this is would need monitoring and regulatory action if necessary.

5.3 Conclusions on cost allocation and tagging

The above arguments suggest that there is merit in re-examining the rules for translating the System Operator's costs into cash out prices. There appears to be a case for incorporating more of the costs of certain critical balancing actions into the cash out prices for those peak periods at which they are primarily directed. At the same time, there is also a case for not incorporating into cash out prices the full cost of actions that are taken primarily with a view to relieving transmission constraints or to meeting other objectives. One suggested rule is that no trades should be included in the calculation of

cash out prices that are taken out of merit order, or at least should be separately identified and justified.²⁹ This suggestion appears to have merit.

The net effect of accommodating both these arguments would need further examination. It might be conjectured that it would lead to lower and less extreme cash out prices in most periods, but to relatively accentuated cash out prices in the more critical peak periods. This of course would need further analysis.

Would such an outcome be a problem? It is not clear that it would. If cash out prices were higher at peak, relative to other periods, this would encourage market participants to take particular care to ensure they were in balance at such times, and would provide additional incentives to generators to provide balancing capacity at peak. If cash out prices were lower off peak, it is not clear that this would be problematic, given that the system is generally long off peak.

It is also worth considering overseas experience. GB cash out prices are reportedly significantly greater, relative to short-term or day ahead market prices, than is the case in other electricity markets.³⁰ This suggests that other markets allocate or attribute a much smaller proportion of the System Operator's total costs to the balancing market than does GB.

This is the case with ERCOT in Texas. For example, zonal congestion costs are directly assigned to those who create congestion. Local congestion costs are part of uplift. None of them are reflected in balancing charges (MCPE). Reg Up and Reg Down are ancillary services that the ISO can procure in day-ahead markets. Suppliers are paid a fee for capacity, then they are price-takers at MCPE in the balancing (BES) market. The capacity fee is included in uplift on market participants as a whole, not in the balancing charges.

It is also noticeable that System Operators in other electricity markets do not seem to experience significant – or significantly greater - problems in balancing their systems. This is despite the lower allocation of their total costs to balancing charges than in GB.

Both these factors suggest that it would be worth reconsidering the allocation of the System Operator's costs to the cash out mechanism. The aim should be to assign costs more clearly, and not to allocate a greater share to balancing charges than could be met from actions in the balancing mechanism.

Removing transmission constraint actions from cash out means that these costs need to be recovered in other ways. A variety of market options is available, including via

²⁹ Cornwall, *Cash-out Review 2007*, section 4.1.

³⁰ “Compared with differentials seen in other electricity markets considered between the real-time mechanism and short-term forward markets, the differentials in price are significant and are almost certainly not representative of real differences in energy costs.” Cornwall, *Cash-out review 2007*, section 4. I understand that there is an order of magnitude difference between GB and elsewhere: in very broad terms cash out prices are about 20 per cent greater than spot market prices in GB compared to about 2 per cent elsewhere. If there really is such a difference, it is not clear why the System Operator is not arbitraging by buying an hour ahead instead of in the balancing mechanism.

transmission rights, nodal pricing, market splitting, uplift, etc. Some of these alternatives offer the prospect of targeting charges more explicitly on those who cause the costs. They do not necessarily preclude generators subject to constraints from participating in the balancing market. There is now much international experience with respect to the identification and recovery of transmission constraint costs.

It ought also to be possible to target other costs more accurately than at present. For example, the System Operator often has to incur considerable cost in order to meet sudden variations in demand and supply within each half hour period. On the demand side, this reflects the beginning and ending of popular TV programmes, dusk and dawn, lighting up time, factory working hours, the switching on and off of night storage heating (though I understand this is now more smoothed), etc. On the supply side these reflect plant falling off the system, failure to meet commitments, network outages, etc. At present a significant part of the resulting costs are recovered in cash out charges paid by those participants that are short or long, rather than charged to those market participants whose actions (or whose customers' actions) necessitated the incurring of these costs. There may be scope to allocate these costs to market participants according to the variability or range of their demand within each period. This in turn would mean, for example, that those suppliers and customers with relatively stable demands would pay less than those with relatively variable demand, and lower costs would be allocated the cash out charges per se. Further study of such possibilities would seem helpful.

6. Ex ante versus ex post pricing and trading

In GB, cash out prices are determined after the end of each settlement period. Some overseas markets set ex post balancing prices too. But others set balancing prices ex ante: before the period begins.

In GB, trading is not allowed after gate closure, which is one hour before each settlement period begins. In some other markets, trading is allowed up to and even into the settlement period.

What are the merits of each arrangement, and is there a case for changing the arrangements in GB?

6.1 Ex post versus ex ante prices

In GB, balancing (cash out) prices are announced 15 minutes after the end of each half hour settlement period. Ex post prices are also used in PJM. But in ERCOT Texas the balancing price is set 10 minutes before each 15 minute settlement period. Ex ante prices are also set in several other US electricity markets.³¹

In general terms, the case for ex ante pricing is to provide earlier information so that market participants can act in a more informed way that is likely to be more efficient.

More specifically, the advantages of ex ante pricing would include the following:

- it would make clearer sooner to the market participants what risks they are running with their current positions, thereby reducing uncertainty
- it would better enable them to do something about it, to the extent that they found this possible
- it would enable those customers buying direct from the balancing mechanism to know in almost real time what prices they were paying
- it would facilitate the development of demand side contracts
- it would provide a benchmark price against which contracts could be written, and could therefore contribute to improved short-term liquidity in the market.

Against this there could be some potential disadvantages. One argument might be that market prices should be based on actual decisions in the market rather than upon a model whose input is a set of expectations of the actions of market participants.

A related concern would be that the information necessary to set balancing prices is not available ex ante. It is true that market participants will have made their bids and offers, which the System Operator will know. And in the light of information provided by market participants, the System Operator can make an informed estimate of total demand and of total generation availability. The former estimates are relatively accurate as the

³¹ I have been told that ex ante prices are used in NYISO, ISO-NE, MISO and possibly CAISO. But there may be some additional subtleties here – for example, in MISO (Midwest ISO), ex ante prices are the basis for the 5-minute dispatch instructions sent to generators, but ex post prices are used for settlement purposes.

period draws close.³² But generation plant can ‘fall off the system’ unexpectedly and there may be generation shortfalls for other reasons.³³ There may also be problems due to network outages. Some losses will be known ahead of time, but not all will be.³⁴ So it may be difficult to forecast the Net Imbalance Volume.

A third concern might be the System Operator’s ability to respond to actions taken by market participants. If the aim is to allow market participants to respond to the earlier provision of information, can the System Operator cope with that in the short timescales now envisaged?

Overseas experience is of some relevance here. Those markets that calculate ex ante balancing prices, such as ERCOT, NYISO, ISO-NE and others, find that close to real time information (about 30 to 45 minutes ahead) is adequate to run the Market Clearing Engine (the system optimisation model) to solve congestion problems and to determine market clearing prices. The amount of inaccuracy due to lack of full information about actual market operations is not materially significant. These markets do not seem to have found that forecasting error is a serious problem. For example, the Independent Market Monitor of the Midwest ISO finds that “the average differences between the ex ante and ex post prices were relatively small”.³⁵ Nor do the System Operators there seem to have experienced problems in coping.

These markets have judged that the benefits that ex ante prices provide, in terms of greater information about balancing prices and reduced market uncertainties, outweigh the possible inaccuracies and occasional ex post corrections. They are continuing with an ex ante approach. I am told that the tendency in the US, at least, is towards rather than away from ex ante pricing.

The basis of calculating cash out or imbalance prices may be important here. If a high proportion of the System Operator’s actions feed into the cash out price, then the latter may be relatively sensitive to the forecasting errors mentioned, so that ex post price might be quite different from ex ante price. But if a significant proportion of those costs is recovered elsewhere, then the impact of the forecasting errors might be less, and the ex ante prices would be more robust.³⁶ It is also possible that the length of the trading period (15 minutes rather than 30 minutes) reduces forecasting error.

³² A rough estimate of the accuracy of the final demand forecast is a mean close to 0 and a standard deviation of approximately 1.5%.

³³ The level of plant loss is quite volatile with the number of generating units tripping off the system on any day ranging from 0 to 12, although the average is around 2 or 3. The average size of a generating unit in GB is about 500 MW. Plant loss represents only a proportion of the generation shortfall experienced on the system, with the other main component being generators producing less output than they indicated they would.

³⁴ A significant proportion of plant loss is instantaneous. Of those losses that are indicated pre-event, the accuracy of the volume involved is quite variable with plant sometimes partially or completely recovering before real time. The ability to accurately reflect the impact on NIV may therefore be quite limited.

³⁵ *2005 State of the Market Report, Midwest ISO*, prepared by Midwest ISO Independent Market Monitor, David B Patton, Potomac Economics, June 2006, slide 90.

³⁶ As noted above, in the ERCOT system Reg Up and Reg Down are ancillary services procured in day ahead markets. If they are called upon, they are price takers for purpose of calculating balancing prices

The potential advantages of ex ante pricing, and the indications that the potential problems are manageable, suggest that this would be a development worth exploring further in GB. It would seem more feasible to move in this direction if, as has been suggested, the calculation of cash out prices excludes the impact of actions taken to deal with transmission constraints.

6.2 Ex post trading

In the UK gas market parties can trade out their imbalance positions for up to 15 days after the end of the month in which the relevant gas day occurs. Would it be feasible to allow something analogous in electricity?

The precise design of such a system would need further clarification. Ex post trading might mean that, after the end of a half-hour period and after cash out prices have been declared, a party that was short in that half hour could trade with a party that was long, such that their net positions would be reduced by the amount of such a trade. This would reduce the exposure of both parties to cash out prices. If there was a single cash out price in that half hour it is not clear that there would be any benefit to the parties in trading. But if there was a dual cash out system, with different prices actually applying in any half hour, then there would be the benefit of arbitrage. That is, a party that was short and facing an SBP of £40 could profitably trade with a party that was long and facing an SSP of £20.

In fact, there would be benefits to trading until one side of the market was eliminated. A dual cash out system would thus seem to be unsustainable if ex post trading of this kind were allowed. The reverse side of the market would be eliminated, and the only price remaining would be the price in the main direction. In effect there would be a single cash out system, although there might be additional distributional issues arising from setting a reverse price in the first place.

Further to the earlier critique of dual cash out prices, would or could the System Operator have acted significantly differently if such arbitrage trades had taken place before gate closure rather than ex post? Or would it have continued to act in the light of the net position of the system as a whole? The argument above is that the cash out price should reflect the net position of the system as a whole.

The System Operator might have a different concern about ex post trading, namely that it could have an effect on the physical aspect of the market. For example, the System Operator might note the positions of the parties at gate closure, and determine upon the appropriate actions to deal with this. But if the parties could trade after gate closure, it would be open to a supplier expecting to be short to contract with a generator that would

(MCPE), but the capacity costs of these services are recovered in uplift, so do not impact directly on the balancing prices. I understand that there are similar arrangements in other US electricity markets, for example NYISO.

not otherwise have chosen to operate. If the System Operator and the short supplier both took action, the effect could be to destabilise rather than stabilise the system.

Although this seems a problem in principle, would it actually be a problem in practice? How many market participants would find it worthwhile to hire the additional staff necessary to carry out ex post trading? Would market participants and the System Operator continue to take actions that were mutually inconsistent leading to unexpected and perverse outcomes? Or would they learn from experience and modify their actions accordingly? It is difficult to see why market participants would not learn in this situation just as they do in other situations. The System Operator would presumably be able to build in expectations about typical responses, in the way that it presently does in other respects. I am not aware that other electricity markets have experienced problems in this respect.

Ex post trading thus needs to be approached with caution. But it seems worth considering as a potentially useful way of extending the ability of market participants to balance efficiently and economically.

7. Innovation, markets, governance and further research

7.1 Innovation and markets

Experience to date suggests a continuing need to monitor and modify the nature of the cash out services provided to market participants. It is not surprising that in an evolving market there should be a continual search for new and better products and services. How best can the System Operator contribute to this process?

Consider, for example, the basis of setting and paying cash out charges. At present this is ex post, on a period-by-period basis. Is there scope for development here?

Previously, a Pool price was set for each half hour. But although suppliers had to purchase on the basis of Pool price (subject to any contracts for differences), they did not insist that all their customers pay on the basis of Pool price. They made this option available to customers, and at one time about 10 per cent of customers availed themselves of this option. Such customers preferred to bear the risks of Pool price fluctuations themselves rather than pay the risk premia that major generators and other suppliers were asking. But at the same time suppliers offered other bases of payment that most customers preferred. Some chose a constant price per unit, others chose risk management terms with a lower unit price but the possibility of supply being curtailed at times of high wholesale prices. Nowadays, many industrial and commercial customers, who are increasingly knowledgeable about markets and price fluctuations, choose prices indexed to fuel prices. This reduces the risk to their suppliers and in turn enables the suppliers to reduce their risk margins and their prices to these customers.

Under the BETTA arrangements, it seems possible that some market participants might prefer to continue to pay the cash out price actually obtaining in each period, as presently determined ex post. But others might like a cash out price set ex ante, just before the period. By extension, some might like a 'fixed price' cash out arrangement, with the cash out price fixed a day ahead or even a month or a year ahead. Some might prefer a single uniform fixed price in all periods, others might prefer specified day-night differentials, yet others might prefer a lower price for non-peak days and a higher price for peak days. The choice of each market participant would presumably depend on its own circumstances and on the costs and risks involved, as reflected in the terms on offer.

What would be the best way to explore and provide such a choice of cash out services? One possibility would be for the System Operator itself to offer a range of balancing services and associated prices. At first sight, this would seem consistent with meeting the needs of market participants.

However, allowing such latitude to the System Operator also has disadvantages. The System Operator has access to aggregate information that other market participants do not. Its credit rating by virtue of its position may put it in a preferable position. Its actions may influence the expectations and actions of other market participants. The System Operator would effectively compete on a favoured basis with market participants that

might offer similar services themselves. For example, potential providers might be financial traders or generators with plant particularly suitable for providing balancing services or short-term response at times of system shortage. I understand that some such market participants have offered cash out products in the past. It would clearly be important for the System Operator not to distort the market, or to provide a product that market participants did not in fact value.

There could be difficulties in distinguishing the costs of the System Operator providing such services from the costs of providing its other services, especially where incentive schemes are involved. There could be problems in ensuring that different options were properly priced vis a vis each other, and did not constitute a favoured or cross-subsidised offering to any particular class of participant. It might be difficult to assess whether any or all of the services provided by a System Operator were financially viable, and should be withdrawn, continued or expanded.

The general suggestion in this paper is that GB arrangements should give more consideration to the approach adopted in many other electricity markets, whereby the System Operator facilitates the operation of markets to provide services, rather than provides the services itself. On that basis, the System Operator should be encouraged to consider what additional services the market participants want, and to promote debate about the means and implications of providing such services, and more generally how to provide greater choice for market participants. The presumption would be that these additional or modified services would be provided via markets, perhaps with a role for the System Operator in facilitating their emergence. This does not necessarily mean that the System Operator has to run the market itself: there are other organisations whose business that is.

In the specific case of the ‘fixed price’ cash out services mentioned above, establishing a more conventional cash out market rather than the present cash out mechanism would seem likely to facilitate the provision of such services by other market participants, if there seemed to be a demand for this. These other market participants would then be better able to trade in the balancing market and to offer hedges against cash out prices of whatever kind the customers were willing to pay for.

7.2 Criteria for System Operator products and services

What should be the criteria for the provision of products and services by the System Operator? A number of underlying principles might be proposed that would not seem inconsistent with criteria presently obtaining in Acts, licences and elsewhere. They might include the following:

- 1) the System Operator should seek (and where appropriate be incentivised) to discover and provide the products and services that best meet the different needs of the market participants, needs that are themselves likely to evolve over time;
- 2) the charges for the services provided should cover the costs of providing those services (subject to the details of any incentive scheme);

- 3) this should also apply for subsets of services and for subsets of customer classes, so that there should be no cross-subsidy between services and market participants; and
- 4) the arrangements should facilitate and not discourage the development of competition and new entry in generation and supply and associated services.

Importantly, consideration should be given to a further criterion:

- 5) wherever possible the System Operator should provide such goods and services indirectly, by facilitating the emergence, provision and operation of markets, rather than directly by its own involvement in the market on its own account or as a market maker.

7.3 Implications for cash out services

In the early days of NETA, there was a suggestion that cash out charges should provide strong signals to encourage market participants to balance their own positions and to discourage reliance on the System Operator. The implication of the present suggestion is in a sense the opposite: the System Operator should be encouraged actively to seek the imbalance business of the market participants. That is, it should seek to provide balancing services that the market participants value, that are better than the services that the market participants can provide for themselves. But, importantly, the System Operator should in the first place seek to do this by helping to establish and operate a cash out market where the market participants can as far as possible trade amongst themselves to deal with their cash out situations.

This approach is consistent with international practice, particularly in the US. There is great reluctance to give undue discretion to the System Operator and a preference for competitive market solutions.³⁷

This is not to suggest that the best outcome is necessarily a greater reliance on the System Operator or even on such a balancing market. Rather, active competition between markets facilitated by the System Operator and other options available to market participants is likely to discover and provide the most efficient and adaptive ways of coping with imbalance and other related issues. Given the increasing importance of new and intermittent technologies, distributed generation and micro-generation, this is likely to become increasingly important.

7.4 Governance

³⁷ Electricity markets are sometimes classified as Min ISO or Max ISO depending on the ISO's level of intervention in the market. For example, the ERCOT ISO is classified as a Min ISO because its main focus is reliability and it has almost no other basis for intervention in the market. PJM and most of the north-east US markets are classified as Max ISO because the ISOs there may also mitigate or adjust market prices and take other actions including enforcement actions as authorized by FERC. But even the PJM ISO is financially neutral and only facilitates market operation rather than acts as a service provider or market maker itself.

The benefits associated with actions of the System Operator will need to be assessed against the costs and risks in any particular situation. This may depend on the type of service or market to be provided. It is not possible here to indicate where or how the boundaries of the System Operator's actions should be set. These are matters to be determined by the governance arrangements pertaining to the System Operator. But perhaps two aspects of these governance arrangements deserve comment.

First, it would seem that good governance requires trust in the System Operator and in the arrangements under which it works. Trust requires an adequately detailed and timely information system, with respect to the activities carried out and their costs and associated revenues. It requires sufficient information and participation to enable accountability. There is scope for improving the present arrangements in these respects.³⁸

Second, if the aim is to provide better markets and services for market participants, then those market participants must play an active role in defining and monitoring the markets and services to be provided. Market participants here should include not only existing generators and suppliers, but also final customers or their representatives as well as traders in the market. They should also include, or take account of, potential new entrants into the market. Clearly the System Operator itself would need to play a vital role in this process.

This is not to say that there is no role for regulation. It needs to ensure that all classes of existing market participants are properly represented or able to participate in the governance process. Regulation also needs to ensure that full account is taken of those interests where representation may not be straightforward, such as smaller customers and potential entrants. It may need to resolve conflicts between all these parties, where conflicting views prevail, or to take a view on matters concerning the public interest or security of supply.

In GB, Ofgem has the last word with respect to modifications to cash out and related arrangements. It has sometimes facilitated consensus and development by signalling the directions or policies that it would favour or not favour. But it has not always seemed predictable. And not infrequently it has rejected Modifications to cash out arrangements recommended to it under the established processes.

In contrast, the PUCT that has jurisdiction over the ERCOT electricity market in Texas established the initial protocols (about 400 pages) in June 2001. It then provided that stakeholders could make whatever modifications they wished, via the ERCOT board, subject to the right of any stakeholder to appeal the ERCOT board's decision to the PUCT. Out of about 500 revisions over the subsequent five years, only about half a dozen were appealed to the PUCT. The PUCT has almost always remained neutral unless a party raised a concern.³⁹

³⁸ See for example the discussion and suggestions in Cornwall, *Cash-out review 2007*, including section 4.3.

³⁹ On the few occasions where the PUCT found a decision not in the public interest, it remitted it back to stakeholders with specific instructions to refine it. PUCT staff appealed a couple of the decisions but PUCT

Subject to all the considerations mentioned, there may be merit in GB regulation seeking to encourage and ensure the discussion and implementation of arrangements agreed among market participants, rather than to impose views of its own about the desirable outcome.⁴⁰

7.5 Further research

At the time when the UK introduced the Pool, there was very limited experience of fully competitive electricity markets elsewhere. And when it moved from the Pool to NETA, there was limited experience about how electricity markets worked that were both competitive and mainly based on bilateral contracts. In each case, the arrangements were inevitably somewhat cautious.

This is no longer the case. Since those times, some markets such as ERCOT in Texas have moved firmly in the direction of competition with bilateral contracts. Other competitive markets have now accommodated contracts to a larger degree than hitherto. In almost all cases there seems to have been a significant development in the use of short-term markets and a move towards 'real time' pricing.

At one time the GB electricity market was in the vanguard of competitive electricity markets. It still is in certain respects. But it seems to be in the rearguard with respect to balancing arrangements. There is now much international experience from which the GB market can learn. This is not least in terms of what it is feasible and desirable to expect the System Operator to do, and in what timescales. The role and limitations of the System Operator as a market facilitator are now well understood. Cost allocation seems to be better developed elsewhere, and markets for balancing-related services are more evident. There is concern internationally about the timely provision of adequate information, and about the effect that System Operator actions can have on market participants. Policies are developing to deal with this.

More extensive and detailed comparative studies of the balancing arrangements in other competitive electricity markets worldwide would provide a better understanding of what kinds of arrangements are available, and feasible in what timescales, and what the effects have been. This in turn would provide a more adequate base for informing decisions about the future directions of cash out arrangements in GB.

has not encouraged that. Staff did question the zonal design of ERCOT and persuaded the PUCT to move to nodal design, scheduled for full implementation by January 2009.

⁴⁰ There is evidence that such 'negotiated settlements' have been successful and welcomed in energy sectors in other countries, notably the US and Canada. Related arrangements are being successfully explored in the UK. For example, the Civil Aviation Authority (CAA) has encouraged a process of 'constructive engagement' between airlines and airports that is proving remarkably productive. For further discussion see Joseph Doucet and Stephen Littlechild, *Negotiated settlements: the development of legal and economic thinking*, *Utilities Policy* 14, December 2006, 266-277.

8. Conclusions

The main conclusions of this paper can be summarised as follows:

- 1) The underlying objective of cash out arrangements is to facilitate competition and efficiency in the electricity market. In the early days of NETA, these arrangements may have been influenced by concerns about generator market power and the ability of the System Operator to balance the system in the time available. Now, these issues are of less concern, and more weight should be attached to the underlying objective.
- 2) The GB approach has emphasised a relatively discretionary role for the System Operator in providing balancing market services itself. It has not facilitated the provision of a balancing market. Where appropriate the System Operator should continue to be incentivised to discover and provide the services that market participants want. However, there would now be increasing value in the System Operator providing cash out services by means of a balancing market rather than by the present relatively discretionary balancing mechanism.
- 3) The arguments for a dual cash out mechanism seem to be associated with concerns about generator market power and about the ability of the System Operator to balance the system. These concerns no longer carry weight. Dual cash out prices distort market decisions towards less efficient arrangements, and are not conducive to accountability of the System Operator or to the development of a liquid short-term contracts market. There would be merit in moving to a balancing market with a single cash out price. This would be more conducive to efficiency and competition.
- 4) Previous debate has focussed on the choice between average and marginal cost pricing. Marginal cost pricing has certain advantages but has been seen as vulnerable to unpredictable prices, a thin market and manipulation of bids and offers. With present dual cash out arrangements and pay-as-bid, marginal cost pricing would drive an additional wedge between total revenues and costs of cash out services. This would not assist the accountability of the System Operator. An alternative scheme such as the use of quantity discounts and premia could secure the advantages of marginal cost pricing without foregoing the advantages of total revenues just covering total costs. But with a single cash out price and without pay-as-bid marginal cost or market clearing pricing need not be inconsistent with cash out services just breaking even.
- 5) Present rules of cost allocation seem to distort cash out prices. The costs of meeting peak demand may not be fully reflected in peak cash out prices. But the costs of actions taken for other reasons than system balancing, such as transmission constraints and intra-period voltage fluctuations, seem to be included when they should not be included, or not included to the present extent. There would be advantage in reconsidering the present rules for cost allocation and tagging. Cash out prices in markets elsewhere are markedly closer to spot market prices than in GB. System Operators in markets with lower differentials do not appear to experience undue problems in balancing the system.

- 6) There would be several advantages in setting balancing prices ex ante rather than ex post, not least to better inform market participants and to provide a benchmark price. The System Operator's potential problems of estimating demand and generation, and the response of market participants, need to be considered. But these problems are evidently not insuperable in other electricity markets, particularly when the costs of transmission constraints are excluded from balancing charges. Allowing ex post trading may have advantages, and it is not clear that the potential difficulties for the System Operator would be a problem in practice. Both these options deserve further consideration.
- 7) There is advantage in continuing to give the System Operator incentives to discover and meet the needs of market participants with respect to innovative balancing services. But where possible this should be by facilitating the provision of markets. Monitoring arrangements need to be developed to improve accountability. Regulation needs to ensure that the process for developing new markets and services adequately reflects the interests of present and future market participants. There would be advantage in regulation encouraging and endorsing agreements reached between market participants with respect to cash out and other services, rather than seeking to impose its own view of the public interest.
- 8) GB has fallen behind with respect to balancing arrangements. There is now much that it can learn from competitive electricity markets elsewhere. More extensive comparative studies of the balancing arrangements in these markets would provide a better understanding of what kinds of arrangements are feasible in what timescales, and what the effects have been, and in what directions policy should usefully develop in GB.

Appendix: Report on Electricity Cash Out Arrangements: Scope of work

The scope of work for this contract will consist of the creation of a report for Wholesale Markets, for publication on our website, no later than 28th February 2007. The report should consider and include the following:

- What should a set of electricity cash out arrangements for the GB market look like? This should consider elements of cash out including the derivation of the imbalance price(s), the role of the system operator, incentives on market participants, how costs are targeted, and any other elements you consider to be relevant.
- You may wish to set out your own views on what the cash out arrangements should set out to achieve, but the following aims should be taken into account:
 - Incentivise efficient balancing
 - Minimise costs to consumers
 - Minimise barriers to entry.
- Where appropriate the report may point to practical obstacles to achieving the most desirable arrangements, but overall should focus on principles.
- If time allows, the discussion should be supplemented with supporting analysis or examples of where the existing cash out arrangements have not provided appropriate signals;
- Also if time allows, draw on international comparisons.