

Response to Ofgem's consultation on electricity cash-out issues

Stephen Littlechild*

23 January 2012

1. Introduction

Ofgem's recent Issues Paper has invited views on whether it should carry out a Significant Code Review that could enable changes to the present electricity cash-out mechanism, and if so whether such a Review should be narrow or broad-ranging.¹

I welcome Ofgem's initiative. The cashout mechanism is indeed fundamental to the functioning of the GB electricity market.

Section 1 of the Issues Paper identifies a number of important developments that will impact on the market generally, and the cashout mechanism in particular. These include the likelihood of an increasing proportion of intermittent renewables on the system, an increasing need for peak and balancing power, the rollout of smart meters and the increased opportunities for demand side participation. In addition, the Government has indicated its preference for introducing a capacity mechanism. It is therefore important to ensure that cashout arrangements are such as enable the electricity market to function in the most competitive and efficient way in the context of such developments.

Section 2 of the Issues Paper identifies a number of issues with present cashout arrangements. I share many of these concerns. It seems to me that a number of the problems follow primarily from the use of a dual cashout price – that is, charging a different cashout price to those market participants that are long compared to those that are short.

The present note examines why a dual cashout mechanism was deemed appropriate in the first place and evaluates the strength of those arguments today. In my view, these arguments are no longer valid. Specifically, there is a stronger economic argument for a single price for imbalances; it is not clear whether dual cashout prices do in fact incentivise market participants to self-balance or in some cases could actively discourage this; and it is not clear that public policy should encourage self-balancing if this is not economic for certain market participants. In addition, it seems likely that dual cashout has introduced a bias in market structure in favour of large vertically integrated entities and against smaller generators and suppliers, and has hindered the development of a spot market reference price and related products that could be more widely traded, thereby increasing liquidity and facilitating new entry.

Section 3 of the Issues paper mentions a move to a single cashout price, and the possibility of introducing an explicit energy balancing market, as possible options to be explored. I welcome exploration of such possibilities.

* Emeritus Professor, University of Birmingham, and Fellow, Judge Business School, University of Cambridge. Without implicating them in the arguments and conclusions of this paper, I should like to thank Nigel Cornwall, Carlos Skerk and Parvis Adib for helpful discussion of this issue.

¹ Ofgem, *Electricity cash-out Issues Paper*, Reference 143/11, 1 November 2011.

I therefore support the proposal that Ofgem should carry out a Significant Code Review that would enable it to explore these issues. A broad-ranging rather than narrow Review, able to look at overseas experience, would seem the most effective way to ensure coverage of the most important options for reform.

2. Background: the design of new trading arrangements

The original Revised Electricity Trading Arrangements (RETA) were proposed by Offer in July 1998, accepted by the Government in October 1998 and set out in Offer's Framework Document of November 1998. They envisaged a Balancing Market for each trading period, with a single imbalance price based on the average cost to the System Operator (SO) of the trades that it needed to carry out in the Balancing Market.

However, the precise operation of the Balancing Market, and the definition of the imbalance price, were not specified at this time.

In July 1999, Ofgem issued proposals for what had become the New Electricity Trading Arrangements (NETA). They referred to a Balancing Mechanism rather than a Balancing Market, and to cashout prices rather than imbalance prices. But there was more than a change of nomenclature. It was now proposed that there be dual cashout prices in each period, whereby those market participants who were short in any period would pay a higher price than would be received by those who were long.

Ofgem's justification of dual cashout prices relied on three main propositions. The first was that this was required in order to reflect the different costs and values of being short and long. The second proposition was that dual cashout would incentivise participants to self-balance. The third proposition was that self-balancing would minimise the role of the SO as 'residual balancer', which was said to be consistent with the then-aims of policy. I now examine these arguments in more detail.

3. The argument for different costs and values

The July 1999 discussion begins with a basic proposition.

"In principle, imbalance cash-out prices should reflect the full costs of imbalances having to be resolved by the SO over relatively short timescales." (p 51)

There is a discussion of the wide range of balancing services, including a variety of flexibility options.

"Calculating the full costs of these flexibility options is very difficult, even if these are confined to the costs incurred by the parties involved, ignoring possible effects on others. Hence, the construction of any pricing system requires an element of judgement." (p. 51)

There is next a discussion of whether those market participants who are short and those who are long impose the same costs or savings on the system, and hence whether they should face the same price. It is argued that those who are short impose

more costs on the system than are saved by those who are long. “This suggests that a dual cash-out price is required.” (p. 51)

The Executive Summary spells out this element of thinking a little more explicitly.

The price paid or charged to ‘out of balance’ market participants varies depending on whether they are over or under contracted. In general, generators who are under contracted (and suppliers who are overcontracted) and ‘spill’ electricity on to the system, potentially imposing balancing costs on the System Operator, will receive a lower price for their electricity than if they had been fully contracted. Suppliers who remain under contracted as the Balancing Mechanism opens (and generators who under-generate), thereby potentially imposing balancing costs, will similarly be charged a higher price than if they had entered into contracts for their full requirements. These charges are reflective of the additional costs incurred by the System Operator in instructing generators, suppliers or customers to vary their output or consumption at short notice to meet unanticipated imbalances via the acceptance of Balancing Mechanism offers and bids. (p 7)

4. An evaluation of the differential cost-value argument

The principle that imbalance prices should reflect imbalance costs seems correct and still appropriate today. It is true that actually calculating these costs is difficult, and judgements have to be made. But there is also an ambiguity about the basis of the calculation that Ofgem skirts around in 1999.

On the one hand, the 1999 discussion suggests an almost moral rather than economic wish to reward those market participants that are fully contracted more highly than those that are under- or over-contracted.² (There is also a suggestion that those participants that are willing to contract later should be rewarded more highly than those that contract earlier.) This implies that the imbalance price for those participants who are long should be lower than the contract price in the market, and that the imbalance price for those participants that are short should be higher than the contract price. But at the same time, Ofgem wants the imbalance prices to reflect the additional costs incurred by the System Operator. Are these two criteria necessarily consistent?

In general, I would say the answer is No. For the most part the SO is not, and cannot be, aware of the evolving contractual positions of the parties, and hence of the extent to which each market participant is or is not in balance at any time, especially for suppliers (retailers). The SO has to deal with the aggregate state of the electricity system as it finds it, reflected in the Net Imbalance Volume in the system. If the SO’s actions are not changed by generator A being 1MW short and generator B being 1MW long, compared to them both being in balance, is there really an economic and cost-related basis for charging generators A and B differently? Surely not.

The impact of individual short and long positions is essentially symmetric (at least, for small changes). Suppose the system as a whole is short. If market participant A is

² E.g. “Participants who spill electricity are not, in any meaningful sense, contributing to balancing the system (except accidentally). Consequently, it seems 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 flexibility costs on the system, withholding these costs from the price they are paid has merits.” (s 5.3.1 p 51)

short, then the SO has to expend resources in the Balancing Market (or Mechanism) to make up for that, and A is charged accordingly. But to the extent that participant B is long, then the SO has to spend correspondingly *fewer* resources to make up for A's short position. More generally, if the market as a whole is short, then a market participant that is long essentially *reduces* the SO's costs, not increases them. It is entirely possible for the imbalance price charged to A to reflect the SO's costs and to be above the contract market price. But the imbalance price paid to B that reflects the reduction in the SO's costs will be the same as that charged to A – that is, *above* the contract market price and not below it.

From an economic perspective, it is not always the case that participants who are spilling electricity receive a lower price for their electricity than if they had been fully contracted, or that participants that are short have to pay a higher price. Market prices are constantly moving. Participants continually decide whether and when to contract depending on their expectations about future prices compared to present prices and upon their attitude to risk. The cost to them – and indeed the cost to the SO - of buying or selling at the last minute may, in the event, turn out to be higher or lower than the cost of buying or selling ahead.

Ofgem 1999 resolved the dilemma by proposing that “buyers of imbalance energy will pay the volume-weighted average accepted offer price, and sellers of imbalance will be paid the volume-weighted average accepted bid price”. This ensured that the buy price would be significantly above the sell price, while at the same time being based on the SO's actual costs. Whether these were the relevant costs is another matter. The cost of a few trades that the SO makes in the reverse direction, in the process of keeping the system as a whole in balance, is not the cost that the SO incurs or saves in dealing with individual market participants whose net position is in the reverse direction.

In justifying its stance, Ofgem's 1999 analysis makes frequent reference to the importance of flexibility. This factor did not feature in Offer's July 1998 proposals. It derives from the Reform of Gas Trading Arrangements (RGTA) programme that was then being implemented by Ofgas and later Ofgem.³ Any flexibility costs incurred by the SO should of course be included in the cost of cashout, to be recovered from market participants. But this does not overcome the fundamental economic point that participants who are out of balance in one direction tend to offset the SO's costs, and should have essentially the same price as market participants that are out of balance in the other direction. In other words, this does not invalidate the case for a single rather than dual cashout price.

5. The incentive to self-balance

³ “The cash-out regime for energy imbalances has also been the subject of considerable debate within RGTA. Ofgem proposed that the price at which out-of-balance shippers are cashed out should reflect both the underlying market prices and the cost to Transco of managing system imbalances. Over the longer term this will be achieved via a two-part cash-out price with a commodity element and a flexibility charge. The commodity element will be based on trades in the OCM [Over the Counter Market]. The flexibility charge will be derived from a daily linepack or storage price, once reliable indicators of the short-term economic value of these services develop.” (Ofgem July 1999 p. 195)

Ofgem's 1999 argument for dual cashout as necessary to reflect costs was backed up by a further argument.

“The use of a dual cashout price regime will incentivise participants to balance their own positions by Gate Closure and hence the actions that the SO has to take should be minimised.” (p 52)

The belief that dual cashout provides additional incentive to balancing is still widespread today.⁴ But is it true? Does the use of a dual cashout price regime incentivise participants to balance their own positions by Gate Closure?

The argument is presumably along the following lines. If a generator expects the system to be short and the main-direction cashout price to be high, it will have a strong incentive to avoid being caught in a short position as a result of under-generating. If, at the same time, the reverse direction price is lower than the main direction price then the generator will have a lower incentive to spill any excess generation onto the system via a long position as a result of over-generating. In this way, the generator is incentivised to balance its position.

Now consider the position of a supplier (retailer) who expects the system to be short but whose own demand is uncertain. It will have a similarly strong incentive to contract ahead to avoid being caught in a short position. But suppose, in the event, that its own demand turns out to be less than it has contracted for, and it finds itself in a long position. Then this surplus cover is worth less than it would have been with a single cashout price, by the amount of the difference between the main and reverse cashout price. In other words, given the uncertain demand and the consequent risk of overcontracting, the dual cashout price seems to *reduce*, not increase, this supplier's incentive to purchase enough cover to balance its position.

Empirical evidence suggests that, in practice, the dual cashout mechanism has *not* in fact led to the parties being in balance, either individually or at the level of the system as a whole. Thus the Issues Paper says that:

“the system is typically short at gate closure in the highest demand periods, but long at other times. This may be evidence of insufficiently strong prices signals at peak. Alternatively, participants with flexible capacity may be electing to make it available via the balancing mechanism rather than selling it in the within-day market.” (fn 12 p 8)

It seems possible that a short system at peak could reflect the point just mentioned. If a supplier is unsure about its future demand, then it has to consider the possibility that it will inadvertently over-purchase cover and have to spill into the system. The prospect of a lower reverse price in a dual cashout system would discourage the supplier from purchasing additional cover that might not be needed. This would particularly be the case in peak periods when demand was unexpectedly high.

⁴ For example, “There are currently different cashout prices for selling and buying electricity. Although this provides a strong incentive for balancing it may not be fully cost-reflective”. *Electricity Market Reform*, DECC consultation document, Cm 7983, HMSO, December 2010, para 9 p 80.

Another possibility is that, in the highest demand periods, which are not easy to predict, participants may simply find it less costly and more efficient to leave the balancing of the system to contracts negotiated by the SO, than to incur the costs of trying to self-balance.

6. Self-balancing as an objective?

Having asserted that the use of a dual cashout price regime would incentivise participants to balance their own positions by Gate Closure and hence minimise the actions that the SO has to take, Ofgem 1999 went on to say:

“Thus, the cash-out prices should also assist in fulfilling the RETA objective of minimising the role of centrally administered mechanisms and facilitating bilateral trading of electricity.” (p 52)

When NETA was first introduced, account had to be taken of the SO’s physical ability to balance the system in real time. Thus, Ofgem elsewhere explained that the incentives to self balance “should serve to limit the scope of the short-term actions that the SO has to take, and thereby make such actions more manageable within short timescales”. (July 1999 p 213)

But where does this alleged “RETA objective” come from? It was not one of the objectives specified in the terms of reference under which Offer produced its initial RETA proposals in July 1998 and under which Ofgem produced its July 1999 proposals. I have not been able to track down any source for it.

It is understandable that minimising the role of the SO might have been an informal objective of policy. Experience in the Pool had shown the limitations of an arrangement in which a central organisation used a set of rules to determine the system price. The thinking on cashout no doubt sought to avoid a return to such an arrangement, and to encourage the bilateral trading that was fundamental to NETA.

However, the Offer July 1998 proposals make no suggestion that parties should be required or even urged to self-balance before entering the balancing market. On the contrary, the balancing market was just one of several opportunities for the parties to buy or sell. It was for each market participants to decide whether, when and how far to contract, depending on the circumstances, preferences and economic situation of each participant.

With the benefit of more than ten years’ experience of NETA (and its successor BETTA), it is not clear that weight should be attached to objectives such as self-balancing or minimising the role of the SO. Bilateral trading and cashout are both now well established, National Grid is an experienced and competent System Operator, and it has been possible significantly to reduce the time allowed for Gate Closure. The more important aim going forward is to enable the market participants to engage in trade and to reduce their risks in the most efficient and effective way possible.

7. The principles of cashout in a possible Review

The Issues Paper proposes four principles by which to assess possible arrangements. In general these seem to me sensible, with one qualification. The first proposed principle says

“Cash out arrangements should, as far as possible, allow and provide incentives for market participants to balance their positions without the need for unilateral actions to be taken by the System Operator.” (para 3.1 p 34)

Certainly cash out arrangements should allow market participants to balance their positions, as far as these participants deem it prudent to do so. The incentives should include prices of balancing services – bought or sold – that reflect the costs of providing these services. It is also generally preferable to allow market participants to trade amongst themselves, making their own decisions in the light of their own judgements, than to rely on a quasi-regulated entity acting on their behalf.

But should participants be particularly incentivised to remove the need for unilateral actions by the SO? Should market participants be “expected to balance their own positions by contracting to buy or sell electricity prior to gate closure”? (Appendix 2 para 1.1 p 23) Experience suggests that most market participants will wish to be largely self-balanced for reasons of risk management. But is there now (if there ever was) a policy objective or operational imperative to require this? And if some types of market participant (e.g. intermittent generators or small retailers) would find self-balancing difficult, should the arrangements seek to shackle them in this way?

There is an alternative perspective. Should we not see the SO as enabling certain kinds of transaction that may be difficult or impossible for market participants themselves to arrange? Participants may wish to be in balance, and they may make some initial contracts in the bilateral market. But in practice some participants have neither the time nor the precise information (about their own supply or demand) to make all the necessary trades as real time approaches. In effect, they ask the SO to make the desired transactions on their behalf. These are of two kinds: they may offer to buy or sell (incs and decs) in the balancing market/mechanism, and they are willing to take the going market price for the remainder of what they take from or put into the market.

Thus, market participants presently have at least three trading options open to them: to contract in one of the bilateral markets ahead of gate closure, to bid into the balancing mechanism (on a pay-as-bid basis) or to take the cashout price. (The Issues Paper raises the possibility of creating a fourth type of option, namely, a balancing and reserve market to replace the second and third of these.) Should we not see the role of the SO (and Ofgem) as to facilitate the operation of all available types of trading, with a view to discovering which types are best suited to each type of market participant, given the costs and risks involved, rather than to minimise the extent to which particular kinds of trading are minimised?

The most efficient operation of the system would mean that each market participant would make its own decision how far to balance its own position in advance of each trading period and how far to use the services provided by the SO. These decisions would be based on the expected costs and risks involved. There should be no presumption that an earlier commitment is more desirable than a later commitment, or

that a bilateral contract with another market participant is preferable to recourse to an indirect contract with another market participant that is effectively brokered by the SO.

This of course requires that the SO has the necessary tools at its disposal to balance the system in real time, and that the relevant costs of the SO's actions are reflected in its charges to market participants. More on this below. It would also be consistent with developing the role of the SO as a facilitator of a Balancing Market, enabling balancing trades between parties. (I wonder whether this might perhaps also include the possibility of cashout option contracts negotiated on a bilateral basis.)⁵

Return now to the four principles proposed in the Issues Paper. Rather than seek to avoid the need for unilateral actions by the SO as in Principle 1, is there not greater merit in designing arrangements to “promote the most efficient operation of the system”, as in Principle 4?

8. Dual cashout and industry structure

The Issues Paper acknowledges an important limitation of dual cashout.

“Dual cashout creates two different prices for the same product in the same period. Standard economic theory suggests that this could lead to sub-economic outcomes. Participants should be able to trade out their positions such that the spread between the buy and sell prices should only reflect transaction costs and risk premia.” (s 2.20 p 11)

In effect, the difference between the dual prices in each period is a tax on a certain kind of trading between market participants. The tax is actually a very high one: on average 27% of price in 2010, and over 50% for 11% of settlement periods.⁶ It is to be expected that a tax of this magnitude would distort the market. The precise manner and direction of this distortion and its effects could benefit from further analysis, but it seems likely to be significant.

One impact could well be upon the structure of the industry. For example, if retailers that are short pay a high price while generators that are long receive a low price, this provides an artificial incentive to vertical integration. As the Issues Paper notes (e.g. para 2.2), this might be expected to disadvantage unduly certain kinds of market participants, particularly smaller and newer participants on one side of the market as compared to larger, established and vertically integrated participants.

9. Is spilling difficult for the SO?

The Issues Paper suggests that the spreads associated with dual cashout

⁵ “[T]he SO contracts forward for some forms of reserve (eg short term operating reserve, or STOR), paying providers an availability fee regardless of when the reserve is used, and a utilisation fee when the contract is exercised.” (Issues Paper s 2.8 p 9) Would it be feasible to enable individual suppliers to contract with individual generators on a similar basis, so that market participants can negotiate their own cashout arrangements?

⁶ Ofgem Issues paper, fn 16 p 12.

“can remove incentives to ‘spill’ into the BM [Balancing Mechanism] rather than trade in the forward or within-day market. If spilling into the BM were to increase, it could cause the SO greater uncertainty and make it very difficult to fulfil its role as residual system balancer.”⁷

The discussion above has questioned the claim that the spreads associated with dual cashout can remove incentives to spill into the Balancing Mechanism. In some circumstances the opposite might be the case. This is an issue that a Significant Code Review might explore more rigorously. It should be possible to provide some economic modelling of the issue.

The claim that, “if spilling into the BM were to increase, it could cause the SO greater uncertainty and make it very difficult to fulfil its role as residual system balancer” also needs to be questioned. If the SO is largely unaware of the extent of any such spilling, why would greater spilling introduce greater uncertainty or greater difficulty into the SO’s world? The Issues Paper notes elsewhere that, apart from peak times, the system as a whole is typically long (i.e. characterised by spilling), and that “this presents the SO with some reserve that is effectively free”. (para 2.15 pp 10, 11) This seems to imply that spilling makes the SO’s job easier rather than more difficult. (The Issues Paper rightly adds “this may not represent the most efficient solution for the system as a whole”, but that is a different point.)

If, at any point, the SO experienced difficulty in fulfilling its role as system balancer, this would presumably be reflected in fewer suitable trades available to it and/or at worse prices. This would immediately be reflected in cashout prices. As in any other market, this would in turn encourage greater supply of the trades that the SO needed and a reduced demand for its residual balancing services. In other words, the market feedback element calls into question any suggestion that abandoning dual cashout would make the SO’s task infeasible.

10. Other concerns about dual cashout

Section 3 of the Issues Paper rightly notes a number of other concerns about cashout prices that are likely associated with dual cashout. These include a lack of transparency and predictability, which may impact adversely on market liquidity, the availability of risk management products, the entry of new market participants and competition generally. Dual prices may place a significant cost on participants that have difficulty in balancing. The dual nature of cashout prices prevents them from being a reliable reference price, further hampering liquidity. Dual cashout also increases the cashflow from energy imbalance charges that is distributed via the Residual Cashflow Reallocation Cashflow (RCRC) mechanism. This introduces an additional element of redistribution between market participants, against those less able to balance.

I agree with these concerns, and have written on them previously.⁸

⁷ Ofgem, Issues paper, s 2.21 p 12.

⁸ “Electricity Cashout Arrangements”, a review carried out for Ofgem, 9 March 2007.

11. Average versus marginal pricing

There are frequently calls for schemes closer to marginal cost pricing that would result in more cost-reflective prices.⁹ The Issues Paper comments that “it is important in principle for prices to reflect accurately the marginal value of energy in each half-hourly period”. (para 2.3 p 8) It notes the difficulty of distinguishing between the costs of balancing the system and the costs of addressing transmission constraints. It says that “This system pollution was one of the reasons why cash-out prices were originally calculated based on the average price of actions taken by the SO rather than the marginal price.” (para 2.4 p 9)

System pollution is indeed a problem, and was indeed a factor adduced in later discussions, but it was not part of the original thinking on the calculation of cashout price. That thinking reflected another difficulty in identifying the relevant marginal cost in the balancing mechanism, which still obtains today.

Offer’s original July 1998 proposal was for imbalance (cashout) prices based on average rather than marginal price. The reasons given were twofold. First, “those who opposed marginal pricing, including customer groups, were concerned that it could increase the scope for gaming the system.” (para 4.48, p 52) This still seems a valid consideration today. Most incumbent market participants have a substantial interest as generators, and therefore as potential providers of imbalance services, as do many potential entrant renewable generators. It is therefore important for Ofgem to ensure that arrangements limit the possibility of gaming and protect the interests of final customers.

Second, Offer questioned the meaning of the term marginal price in this context.

“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 market clearing price throughout the period.... it is not obvious that imbalance prices set out on a marginal basis would be more efficient than those set equal to average cost.” (para 4.49 p 52)

Although gate closure has since been brought much closer to the trading period, this remains a valid point.

Suppose that all the trades relevant to balancing a particular half hour trading period were selected by the SO at one point in time, and were all of the same character, and were all relevant only to balancing rather than to other system conditions. Then the SO could select the trades in order of price. If the SO expected the level of imbalance to be slightly higher or lower, the SO would seek to purchase or sell a slightly greater or lesser extent of the highest value trade. The highest value trade would represent the SO’s marginal cost.

⁹ E.g. “The current scheme is ‘pay-as-bid’ and the imbalance price is the average of the most expensive 500MWh of balancing actions. A scheme closer to marginal pricing would result in higher and more cost-reflective prices at times.” *Electricity Market Reform*, DECC consultation document, Cm 7983, HMSO, December 2010, para 9 p 80.

In practice, however, the situation is more complex. In order to balance the system in a particular half hour trading period, the SO makes numerous trades in that period and during at least two earlier periods. The SO's actions are determined not only by the trades on offer to it at any point in time, but also by its evolving expectations about the likely extent of imbalance, and the trades expected to be available to it at a later period. The SO's marginal cost will thus depend on *when* it revises its expectations, and what trades it expects to be available to it at that time or later. The highest price trade over the whole period is not necessarily the one that the SO would or could have engaged in to a greater or lesser extent in order to address a higher or lower expected imbalance. Moreover, the relevant marginal cost for the trading period is presumably not the cost associated with a particular decision actually taken by the SO, but the cost that would have been incurred or saved if the SO had held a different expectation about imbalance throughout the whole three or more periods.

Because there was “no obvious definition of marginal cost” in these circumstances, Offer in 1998 concluded that “it is not obvious that imbalance prices set out on a marginal basis would be more efficient than those set equal to marginal cost”. (para 4.49 p52) Ofgem 1999 endorsed this view.¹⁰

Since then there have been various modifications in the calculation of cashout price, generally seeking to identify more accurately the marginal price given the practical realities just mentioned. There is no reason why such refinements should not continue to be explored, with a view to better identifying the trades that are actually at the margin (with respect to balancing) within each set of purchases. But the more that judgement has to be applied, the less transparent and predictable is the procedure of setting cashout prices. And while methods that produce higher cashout prices are likely to appeal to generators and integrated companies, they are correspondingly less likely to appeal to small suppliers and customers.

12. A Balancing Market?

Minimising the need for subjective judgements is one of the potential appeals of creating an explicit balancing market, such as the one to which the Issues Paper refers.¹¹ Looking at my earlier paper again, it seems to me that it contains some ambiguities (e.g. as to the actions taken by the SO in that market) and also raises some questions (e.g. as to the incentives on the parties). There are also two main elements, which could usefully be considered separately.

First, there is the possibility of a balancing market to be held just before each trading period commences. Provided there is a demand for such a market (in addition to the recently established N2EX day-ahead auction market), I see several advantages here, including a) to facilitate any remaining trades that might be too time-consuming for the parties to carry out on a bilateral basis or over-the-counter, b) to provide a vehicle

¹⁰ E.g. “Given that Gate Closure is four hours ahead of real time ... there is a significant risk that marginal prices would be set by unrepresentative actions (for example, a high price offer accepted early in the trading window).” (Ofgem 1999 p 79)

¹¹ Issues Paper fn 23 p 18 refers to my paper entitled “A proposal for a balancing market to determined cashout prices” (17 April 2007). This was a specific suggestion that followed on from discussion of my broader review of “Electricity Cashout Arrangements” (9 March 2007).

for the SO to make balancing trades, and c) to reduce the extent to which the SO has to act in real time to balance the system. Such a market would provide a final spot market price and facilitate competition. It might be particularly useful for smaller and non-integrated players, and would provide additional public market information that would be conducive to liquidity and the development of secondary markets and other products.

Separate from this is the question whether this ex ante price should subsequently be used as the cash-out price for remaining imbalances, and if so whether the balancing market price should be adjusted by using the SO's prediction of the Net Imbalance Volume. A case for this was set out in my note: it would provide an easier and cleaner way to determine a market-based cashout price, and it would give clear signals. However, I now wonder about certain questions. Would it provide an incentive to distort prices by withholding demand or supply in the preceding balancing market? How far is it possible and desirable to distinguish between the SO's actions and predictions? Would this ex ante spot price be a sufficiently accurate reflection of the ex post costs subsequently incurred by the SO in actually balancing the system, or would it introduce distortions of its own?

We also have to realise that the world has moved on since the last discussion. Amongst other things, there have been significant moves towards real-time balancing and pricing that could not have been envisaged at the time that NETA was designed, or even when it was last modified. For example, the ERCOT system in Texas, like the UK system, is based on bilateral trading, which accounts for more than 90% of the electricity used there. ERCOT administers a Day-Ahead market that allows some predicted imbalances to be traded. On 1 December 2010 it introduced a nodal Locational Marginal Pricing system that involves real-time balancing market auctions for each of more than 4000 nodes for every five minute interval (taking place one hour ahead of each operating period). Each auction determines the price that obtains for energy imbalances – positive or negative – at that node during that five minute period. I am told that more than 98% of the time the balancing price is close to the earlier market price, but there are significant differences when there are spikes in demand or supply.

The Texas system is not immediately applicable to the UK. These are nodal rather than zonal prices, and the ERCOT arrangements provide for a little more commitment by the parties and more scheduling by the SO than is at present the case in the UK. The ERCOT system has not found it necessary or desirable to introduce a capacity mechanism. Nonetheless, there is scope to learn from this and other overseas systems. A broad-ranging Significant Code Review should therefore provide for this.

13. Conclusions

There is a strong case for reviewing certain aspects of the present electricity cashout arrangements. This note has made four main points.

- Whatever the merits of a dual cashout mechanism in facilitating the introduction of NETA, the economic and practical case for it is no longer compelling. It may well have provided an artificial incentive to vertical integration, favouring incumbent competitors over entrants, and hindering the development of more liquid traded markets.

- In specifying principles governing such a Review, it is not clear that the arrangements should seek to minimise the role of the SO, or to “provide incentives for market participants to balance their positions without the need for unilateral actions to be taken by the SO” (Principle 1). It seems more appropriate to identify and implement cashout arrangements to “promote the most efficient operation of the system” (Principle 4).
- Although there is scope to refine the calculation of cashout costs, suggestions for higher cashout prices at times of greatest pressure on system capacity, and for use of marginal cost pricing, need careful consideration. The original case for cashout pricing based on average cost rather than marginal cost was that marginal cost was not necessarily equal to the highest cost incurred by the SO and that marginal or highest cost offered greater scope for gaming the system. These arguments remain valid today.
- The case for a balancing market rather than mechanism deserves further investigation. Whether such an ex ante balancing market should also be used for ex post setting of balancing charges needs further consideration. The latest international experience will be of great relevance here.

The conditions of the future electricity system will be different from those today in numerous respects – for example, less flexible renewable energy and more responsive demand side participation. The Government has also indicated a wish to introduce a capacity mechanism. Energy balancing and related pricing are likely to become increasingly important. This makes it all the more necessary to put the fundamentals of the cashout system on a more efficient economic basis. This would be consistent with moving further towards a balancing market rather than a balancing mechanism.