Promoting Efficiency and Security in the NETA Pricing Arrangements

Barclays Capital

July 2003

1. Introduction

Events of last winter have raised many questions as to the effective functioning of the electricity market in England and Wales. In particular, short-term and forward energy markets have largely failed to respond effectively at times when the system has been under particular stress. The most striking example of this was on 10 December last year when several events meant that the system was very short of energy and associated reserves. Despite this, prices in short-term markets on 10 December failed to respond. Indeed short-term prices on 10 December were lower than on the day before and the day after, where the supply-demand balance was significantly less tight. Looking forward, the failure of short-term prices to respond to market shortages raises longer-term concerns about the ability of the forward markets to respond effectively to emerging shortages and consequent risks to security of supply. This concern is borne out by recent experience in the forward market for next winter, where despite repeated warnings from NGC about the likely lack of sufficient operating margin, there has been little movement in forward prices or increases in projected generation availability to address the shortfall.

The increased likelihood of demand interruptions in the coming winter means that there is an urgent need to review the cash-out and balancing arrangements in England and Wales to ensure that they are robust to generation shortages. In this paper, we review the current arrangements and make several suggestions for change which we believe will lead to improved market price signals of emerging generation shortages.

To review the current arrangements effectively, it is important that problems are identified and solutions proposed within a robust overall theoretical framework. To this end, in section 2 below we present a simplified "model" for how competitive electricity market should work and analyse several fundamental market failures in electricity that must be accounted for in the design of an efficient electricity market. Section 3 then uses the backdrop of an "ideal" market design to analyse the current England and Wales arrangements. Based on our analysis of the current problems in England and Wales, Section 4 develops some proposed developments to the balancing and settlement arrangements to ensure more efficient pricing of imbalances, improved treatment of balancing services and greater clarity on the overall reliability policy.

2. Electricity Market Pricing

Before assessing the current arrangements in England and Wales, it is important to have a benchmark "model" of how an efficient electricity market should be structured. As we describe in section 2.1 below, at a very basic level, electricity can be treated like any other commodity with prices being driven by supply and demand fundamentals. However, in practice, implementing a successful electricity market requires moving beyond the simple competitive orthodoxy to address several special features of electricity markets, including the delivery of power over a network, the need for centralised real-time balancing of supply and demand and the absence of demand response to price signals. Economically, these features can be characterised as "market failures" which require deviation from, and amendment to, the standard competitive model in designing an efficient electricity market. We explore the source of these market failures and the associated policy response to correct these failures further in section 2.2 below.

2.1 The Basic Competitive Paradigm

As with any other commodity, electricity market prices should be formed by the competitive interaction of supply and demand such that for each delivery period:

- the market price should equal the marginal cost of production when supply is sufficient to meet demand; and
- the market price should rise to equal the marginal value of curtailing load when all available supplies have been exhausted.

In this competitive paradigm, spot market prices are sufficient to balance supply and demand and thereby to ensure an efficient pattern of production and consumption over time. (In economic terms, this is known as "allocative" efficiency.) Moreover, competition to produce and consume ensures:

- productive efficiency, ie, generation is produced in least cost as producers compete to provide generation at or below the market prices; and
- consumptive efficiency, ie, those consumers who value the product at or above the market price are served.

Over time, competitive market prices will drive efficient maintenance and investment decisions. When there is insufficient generation capacity, prices will rise to the marginal value of reducing consumption to "ration" demand to the available production. This provides "scarcity rents" (ie, the excess of prices over production costs) to generators. These rents provide revenues that contribute to generators' fixed operating costs, capital costs and financing costs.

Looking forward, the expected value of these rents drives efficient investment in generation. When there is insufficient generating capacity, generators will expect prices to rise more frequently to rationing levels. This increase in the expected scarcity rents increases the expected value of generating capacity – compared to the fixed costs of providing additional capacity – thereby signalling the need for generators to build new capacity and to compete away this value until it is just sufficient to cover the cost of additional generation capacity.

Similarly, if there is an excess of generating capacity, expected revenues will be insufficient to cover the fixed capacity costs of generation and generation will be mothballed or closed thereby increasing expected revenues to a level just sufficient to cover the costs of generation capacity.

2.2 Special Features of Competitive Electricity Markets

The above discussion of competitive market pricing in electricity markets has underlined the liberalisation of electricity markets globally. However, it presents a misleadingly simple view of how competition in electricity market could and should work. In particular, electricity markets exhibit the following three main market "failures":

- Incomplete demand side response to market signals;
- The need to coordinate dispatch in the short-term to balance supply and demand; and
- The requirement for a centralised system of imbalance settlement.

Each of these has implications for the design of an electricity market, which takes us away from the pure competitive paradigm described above. We discuss each of these failures and the associated policy responses further in the following sub-sections.

2.2.1 Demand Side Market Failures

The simplified paradigm above assumes that prices rise to the marginal value of demand when generating capacity is short. This assumption assumes that demand sees the price signal in each delivery period and can respond to that price signal by reducing consumption. However, in practice, demand-side signals are seriously constrained by metering technology. Large industrial and commercial customers are metered half-hourly which means that they – or their supplier – will see and have the opportunity to respond to emerging generation shortages. However, the half-hourly consumption of domestic and small commercial customers can only be estimated with aggregated profiles and longer-term (eg, annual) meter readings. In practice, this means that a significant portion of demand cannot and does not respond to market price signals in an individual half-hour. The absence of demand-side response leaves a gap in the competitive determination of electricity market prices. In particular:

- prices at times of shortage do not necessarily reflect the marginal value of electricity to consumers. While prices may nevertheless rise above marginal costs at these times, eg, due to exercise of market power by generators, there is no guarantee that this will send an *efficient* signal on the value of generating capacity to the market. Price rises due to the exercise of market power also carry a risk of regulatory intervention to ameliorate these prices.
- Demand must be disconnected to match demand to available capacity and it is impossible to target those disconnections at those customers belonging to suppliers that have failed to procure sufficient capacity.

Reliability of supply therefore takes on the characteristics of a public good since individual suppliers - and their consumers – who have failed to procure sufficient electricity, cannot be excluded from its consumption at times of shortage. Uncorrected this would lead to the universal under-provision of generating capacity as suppliers attempt to "free ride" on others' purchases of generation capacity. To correct this problem, electricity market designs usually incorporate a reliability policy along one of the following lines:[1]

- Requirements for suppliers to procure a set proportion of installed capacity. (This forms the basis for many of the North-Eastern US markets and is a key component of FERC's standard market design).
- Value of lost load (VOLL) pricing such that market prices rise to a regulated proxy for the marginal value of consumption in the event that demand is disconnected. This approach is adopted in the Australian markets and variants based on combining the value of lost load with the likelihood of load being lost have been used elsewhere (eg, in Spain, the UK Pool and some South American markets).
- Operating Reserve Pricing such that the price rises to a regulated price when the system runs short of operating reserves. (The PJM market design and FERC's standard market design include this feature).

All approaches recognise the absence of universal demand response in the setting of electricity prices and therefore substitute a proxy to ensure that market participants attribute a value to generating capacity. Each potential system has their pros and cons and we will not discuss these at length in this paper. However, it is clear that pending technological developments to improve demand response, all electricity markets require a reliability policy to address the market failure resulting from the absence of real-time demand response.

2.2.2 Short-term coordination of dispatch

In electricity markets, the system operator needs to take central control of the dispatch of production (and consumption) at some point before delivery to ensure that supply matches demand in real time. This allows for production to be matched to consumption while respecting several operational constraints, eg:

- The need to maintain a stable frequency by matching supply and demand at all times;
- Voltage and thermal constraints on the flows of electricity across an interconnected system; and
- Dynamic constraints on the output from a range of different production technologies.

The system operator will typically operate a "balancing mechanism" to achieve this role, by

[1] An excellent discussion about the relative merits of different reliability policies is contained in Part 2 of "Power System Economics" by Steven Stoft, IEEE, 2002.

calling on generation (and some demand) to provide more or less energy and to provide regulating reserve or other services ancillary to the basic delivery of energy (eg, the provision of reactive power, fast response, frequency response etc). The system operator's objective is typically to procure the required energy and associated services at least cost – subject to these constraints. This obligation ensures efficient production and consumption and mimics the result that a full market in these services would achieve (if such a market was feasible).

2.2.3 Imbalance Settlement

Coupled with the requirement for short-term balancing, electricity is *automatically delivered to consumers* via the transmission system. This differentiates electricity (and gas) from other commodity markets, where delivery is an integral part of the forward markets. Automatic delivery raises a specific economic problem of "free riding", ie, consumers may take more (or less) than their suppliers have contracted for or that generators deliver more (or less) than they have sold under contract. The possibility of free riding means that a centralised system to calculate, price and settle imbalances between contracted and actual deliveries is required. The price for settling imbalances takes on a crucial role in electricity markets, since the imbalance price represents the opportunity cost of contracting for power in advance of delivery, ie, a failure to contract will incur the imbalance cash-out price. The price at which power will trade in forward energy markets will ultimately reflect the price for imbalances (ie, failing to contract) because of the ability to arbitrage between forward and imbalance settlement by contracting for more or less than one's likely physical deliveries.

Given arbitrage between forward markets and imbalance cash-out, an efficient market design requires imbalance prices to reflect the marginal cost (or marginal value) of the energy actions taken by the system operator in balancing the system. (Since electricity is traded as a simple MWh commodity in each delivery period, it is important that the cash-out price, as far as possible, should reflect the underlying marginal cost/value of delivering a MWh in each delivery period.) Given the many non-energy operational constraints on balancing the system, setting an efficient energy-only cash-out price requires account to be taken of:

- Temporal externalities links between different market periods, eg, a generating unit committed for one half-hour may also need to be run for several subsequent periods;
- Network externalities which drive differences in the market value of electricity at different points on the system;
- The pricing of products jointly produced with energy (eg, reserve and reactive power); and
- Effective public goods such as frequency response where the marginal cost of consumption usage from any individual market participant cannot be measured.

Electricity markets therefore incorporate many techniques to isolate the impact of these operational constraints and to distil out the marginal cost/value of the underlying energy.

Techniques to isolate the marginal energy price therefore include:

- Explicitly procuring and costing non-energy products (eg, reserves) outside of the energy market;
- Using hypothetical schedules to effectively "tag out" the effect of decisions taken for locational reasons or to meet obligations outside of the delivery period;
- Explicitly costing locational influence, eg, via nodal pricing or explicit definition of transmission rights.

In practice, there is no overriding consensus on the best way to calculate the energy-balancing price and many different approaches have been taken. Whatever technique is adopted, the underlying objective is clear - the cash-out price must reflect the marginal cost/value of matching demand and supply for MWh delivered over the delivery period and, wherever possible, should exclude the effect of non-energy constraints or products.

2.3 Summary: Rules for Electricity Market Design

The above discussion has highlighted several fundamental features of electricity markets which move electricity away from a simplified competitive paradigm based on the interaction of supply and demand to set real-time power prices. Each market failure demands a policy response to ensure that – as far as reasonably possible – the balancing and settlement arrangements underwrite competition in forward markets for the basic "electricity commodity". The policy responses resolve to four basic rules for electricity market design:

- Balancing actions should be taken in least-cost order – subject to operational and locational constraints – to ensure efficient production and consumption decisions
- Cash-out prices should reflect the marginal cost (or value) of matching demand and supply for energy only and, wherever possible, should exclude the impact of non-energy actions (eg, reserve procurement) and impact of operational and locational constraints.
- There should be efficient arbitrage between cash-out prices and forward markets.
- A proxy is required for the marginal value of demand at times when flexible demand response has been exhausted;

In section 3, we review the current arrangements in England and Wales and highlight those areas where the current arrangements fail to meet these market design rules. Section 4 then proposes some modifications to the current framework which should produce a more comprehensive and efficient set of arrangements.

3. Assessment of the E&W Electricity Market

3.1 There is No Clear Reliability Policy

NETA removed many elements of the reliability policy present under the Pool, ie:

- The Pool's system of capacity payments based on LOLP and VOLL;
- A licence requirement on suppliers to purchase electricity in the Pool up to the value of VOLL;
- An obligation on suppliers to meet a generation security standard of supplies being discontinued in no more than 9 years in any period of 100 years;[2]

While these changes were well-motivated and intended to remove the significant manipulation of capacity payments by generators under the Pool, the introduction of NETA effectively saw the removal of any clear reliability policy. While remnants of a policy remain in the new licence conditions and the Balancing and Settlement Code, the implication of these conditions for security remain unclear. In particular:

- Suppliers retain an obligation to meet the "reasonable demands" of domestic electricity consumers.
- NGC has statutory obligations "to develop and maintain an efficient, co-ordinated and economical system of electricity transmission" and "to facilitate competition in the supply and generation of electricity".
- While the Balancing and Settlement Code provides for demand side response, there is no longer any identifiable proxy for the absence of demand response from non-half hourly customers either there or in the wider regulatory framework.
- Standard Licence Condition C2 of NGC's licence prohibits NGC from purchasing or acquiring electricity for sale or other disposal to third parties except with the consent of the Authority.
- Special Condition AA3 of NGC's licence prohibits it from purchasing or otherwise acquiring electricity except pursuant to the procurement of balancing services.

In practice, the reliability policy in England and Wales is a function of how NGC and Ofgem interpret these provisions. For example:

- It is not clear how suppliers satisfy Ofgem that they have met the requirement to meet the reasonable demands of domestic consumers. It is unlikely that this requirement can be interpreted as "meeting all demand and at all times", but that begs the question of what level of security suppliers should procure to meet the obligation.
- A wide interpretation of NGC's statutory obligations could include a requirement to introduce a reliability policy under the Balancing and Settlement Code since this is

[2] This condition remains in the licence and applies to the supply of customers in Scotland. Given that the Scottish wholesale price is currently based on the prices in England and Wales, it seems anomalous to retain this condition in Scotland but not in England and Wales.

required to “facilitate competition in the generation and supply of electricity”. However, the Balancing and Settlement Code currently includes no explicit recognition of the absence of demand-side response. Reliability therefore becomes a function of how high cash-out prices might rise at times of shortage, which in turn relies on the extent to which generator offers rise to rationing levels at times of shortage. However, in the absence of a proxy for demand-side response (and with inadequate arbitrage between the balancing mechanism and forward markets) there is no guarantee that this process will produce an optimal level of generating capacity and reliability. For example, NGC’s balancing principles statement appears to require it to accept all offers in the balancing mechanism even if these rose to several hundreds of thousands of pounds, which comfortably exceeds the likely value of lost load. An additional concern is that the current arrangements appear to rely implicitly on the exercise of generator market power at times of shortage that raises additional risks and uncertainties related to regulatory intervention to correct this.

- The definition of “balancing services” in supplementary standard condition C1 includes “other services available to the licensee in operating the licensee’s transmission system”. This is a very broad definition, which does not necessarily limit the services to those required “for the purpose of balancing the licensee’s transmission system”. This latter qualification features in the definition of “balancing services activity” but not in the definition of “balancing services”. However, condition AA3 uses the broad definition in its prohibition of electricity purchases other than the “procurement or use of balancing services”. In addition, supplementary standard condition C2 includes a blanket prohibition on purchasing or otherwise acquiring electricity except “with the consent of the Authority”. Taken together, these conditions create some confusion on whether NGC will procure generating capacity at times of shortage and whether Ofgem will permit it to do so. Specifically, it is not clear whether NGC has to balance the system with the capacity that is made available by generators or whether NGC can go further than this and procure additional generating capacity in advance of an expected shortage as a “balancing service”. (This would appear consistent with a wide definition, but not necessarily with the narrower definition.) The Drax contract entered into by NGC in November 2002 indicates that NGC does indeed have the flexibility to procure generating capacity in advance of an expected shortage. However, this apparently conflicts with the general prohibition on NGC from purchasing or acquiring electricity for sale or other disposal to third parties. It also conflicts with prevailing opinion – as informed by NGC, Ofgem and DTI statements - that the scope to intervene in the market is – and should be - strictly limited.

In addition to these uncertainties, several genuine gaps remain in the overall reliability policy. Suppliers have no explicit obligation to meet the demands of non-domestic consumers. Such an obligation is arguably not required for half-hourly customers, since suppliers could theoretically enter contracts with these customers to procure a demand reduction in the event of shortages (and the attendant high prices). However, the extent of such contracts is unclear and *in practice,* there is likely to be limited demand response even from half-hourly metered customers. (Indeed, to the extent that demand response is contracted for, this appears to be directly with NGC rather than with suppliers, which indicates that it is not being directly reflected in the market.)

The obligation to meet the reasonable demands of domestic consumers also fails to account for supplies to non-half-hourly metered commercial and industrial consumers, which leaves a genuine question on suppliers' responsibility for continuing supplies to these sites. This lack of clarity on the demand side is compounded by problems with suppliers' incentives to procure additional generating capacity in advance when disconnections become likely. In the event of demand disconnections, individual consumers cannot be isolated for disconnection should their supplier have failed to procure sufficient generation. The result is that consumers are disconnected en masse on a non-discriminatory basis. This has the perverse result that some suppliers in affected areas will have lower deemed meter readings and will effectively be spilling into cash-out and actually get paid for any of their customers' load that is not served. This provides a perverse incentive for suppliers to continue procuring capacity when the system is likely to be very short which exacerbates the weak signal currently provided by the cash-out price arrangements.

In summary, there is currently no transparent and explicit reliability policy to address demand-side market failures. While market and regulatory responses may allow NGC to underwrite security, the circumstances in which it might intervene – or is permitted to intervene - remain unclear. In turn, this uncertainty is likely to distort the market's response to emerging shortages and forward market prices are unlikely to signal the true value of generating capacity.

3.2 Cash-out Prices Fail to Reflect Underlying Marginal Costs of Balancing

The analysis in section 2.3 argued that cash-out prices should reflect the marginal cost (or value) of matching demand and supply. Under NETA, offers and bids in the balancing mechanism are accepted on a pay-as-bid basis, with the accepted offers being averaged (with appropriate weights) to calculate the main cash-out price, ie, System Buy Price (SBP) when the system is short and the System Sell Price SSP when the system is long.

The combination of pay-as-bid acceptances and weighted average prices was introduced to counteract concerns about gaming of marginal prices by generators in the Pool. In *theory*, this rule should produce the same results as marginal pricing for acceptances and in setting the cash-out price. In determining offers and bids into the balancing mechanism, market participants should set prices based on their expectation of the marginal energy offer (bid) to be accepted. With perfect information and foresight, the weighted average of offers should therefore tend to the marginal accepted offer. In practice, however, experience with the balancing mechanism is that the theoretical assumption of rational predictions of the marginal accepted offer is rarely fulfilled and, in particular:

- A wide-range of offer prices typically exists. In part, this can be attributed to the "dual" function of the balancing mechanism in procuring both system and energy actions. A plant providing locational energy or other balancing services may therefore bid based on their expectation of the value of these related services in addition to the underlying value of energy.
- Cash-out prices persistently underestimate the marginal cost of balancing supply and demand at times of shortage.

The consequence of this is that cash-out prices (and hence forward market prices) will fail to reflect the true underlying costs of shortage. In addition, dispatch will be inefficient if market participants mis-estimate the likely marginal acceptance price. For example, plants with low marginal costs may set offer prices above marginal cost in the expectation that NGC will accept higher prices from more costly generation in the BM. However, if they overestimate the marginal acceptance price, NGC will not accept their offers and other plants with higher marginal costs, but more keenly priced offers, will be accepted. The net result will be higher, inefficient costs of generation, although NGC has continued to dispatch "efficiently" according to the offered prices. The efficiency loss due to these problems appears relatively small when the system is long; accepted bids and SSP appear to broadly reflect the marginal cost of backing units down. However, these problems can be catastrophic at times of generation shortage, where prices spanning thousands of pounds lead to distorted price signals and likely dispatch inefficiencies. This range of prices was demonstrated clearly on 10 December 2002, where NGC accepted offers up to £10,000 MWh for periods 35 and 36, but where SBP was £270/MWh and £261/MWh respectively.

There is therefore strong evidence to suggest that the current system of pricing under NETA is likely to underestimate the marginal cost and value of electricity at times of relative scarcity. However, it is at times of shortage where the accuracy of pricing signals becomes most important if the market is to receive an adequate signal of the value of generating capacity.

3.3 Cash-Out Prices Do Not Adequately Capture Standing Reserve

In other commodity markets, market participants write options against the possibility of high spot prices, which in turn depend on the marginal cost of production or, in times of shortage, the marginal value of reducing demand. In power, there is no pure real-time spot market price since in addition to the balancing mechanism, NGC uses pre-contracted reserve "options" to match supply and demand in real-time.[3] The use of pre-contracted options obscures the underlying "true" marginal cost of balancing that would need to be reflected in market prices

---

[3] In this paper, for simplicity our analysis focuses on the use of "standing" reserves procured by NGC. However, in so doing we are effectively using "standing reserve" as shorthand to refer to any pre-contracted firm reserve purchases made by NGC irrespective of whether they fall strictly into the definition of standing reserve. The purchase of Firm Regulating Reserve contracts would therefore also be included in our definition. The treatment of "non-firm" operating reserve, eg, that procured via the balancing mechanism would fall outside the scope of this definition, since the option fees for procuring these services are implicit in the balancing mechanism bids/offers and undo bids/offers.

if the plants providing standing reserve were only paid when NGC calls them to generate. Specifically, prices would need to rise high enough to cover the fixed and variable costs of standing reserve providers in those periods in which they were used. Pre-contracting for these reserves smoothes the revenues available to these providers, secures their availability on an annual basis and reduces the risk premia that they would need to extract for highly uncertain running regimes. However, in addition to smoothing the payments to standing reserve providers, the current rules also effectively smooth the reflection of those payments in cash-out prices. This distorts the cash-out price signals sent to the market when generation is short and NGC uses reserve in two main ways:

- The option fees paid to standing reserve generators feed through into cash-out prices via the Buy Price Adjustment as the average of the capacity fees paid for those periods in which they are made available to provide the service (ie, roughly half of the hours in the year) rather than in those periods in which they actually produce.
- The standing reserve tenders require providers to offer a capacity fee (MW) and utilisation fee (£/MWh). NGC's then assess the tenders based on likely usage of the reserve and hence the likely average cost being paid when capacity and utilisation fees are combined.[4] The result is often utilisation prices that fail to reflect the underlying marginal costs of production.

In addition to the generic problem of capturing standing reserve correctly in cash-out prices, there is also a particular problem with the provision of standing reserve from non-BM participants. NGC currently has over 500 MW of standing reserve procured from non-balancing mechanism sources.[5] While this reserve will therefore contribute to system balance, the costs of using this reserve do not feed through into cash-out prices at all. This is an additional source of distortion to price signals to the market at times of capacity shortage.

In summary, the current methods for accounting for standing reserve in cash-out prices understate the full marginal cost of those reserves at times of capacity shortage. The result is lower cash-out prices than would otherwise obtain and hence lower prices for generating capacity that is not explicitly procured via the standing reserve tenders. The result is likely to be inefficiently high levels of closure for relatively high marginal cost generation that sits in the generation stack just below that capacity contracted to provide standing reserve.

---

[4] See "National Grid Standing Reserve Market Report for Contracts Effective from 1 April 2002 to 1 April 2003, NGC, September 2002: http://www.nationalgrid.com/uk/indinfo/balancing/pdfs/2009_Standing_Reserve_Report_re_2002_2003.pdf

[5] **"Standing Reserve Tender Final Results Statement for Agreements Commencing 1 April 2003" available on NGC's website:**
: http://www.nationalgrid.com/uk/indinfo/balancing/pdfs/Final_Results_Statement_01042003.pdf

3.4 Short-term market prices fail to reflect cash-out prices

The previous sub-sections have highlighted several reasons why cash-out prices fail to reflect the underlying marginal costs of balancing. In particular, they have demonstrated that the result is likely to be cash-out prices at times of shortage are likely to be *lower* than they should be to reflect the true costs of a shortage of generation capacity. However, a cash-out price calculation that reflects marginal costs and values is only one of the necessary conditions for an efficient market signal of the value of generating capacity. The other requirements are that there should be efficient arbitrage between cash-out prices and markets in advance of gate closure and that the impact of constraints and non-energy related actions (eg, reserve procurement) should be removed from the determination of energy cash-out and market prices.

One of the main concerns stemming from the events of 10th December 2002, is that not only did cash-out prices fail to rise sufficiently high to reflect the imminent demand disconnections, but that short-term exchanges prices also failed to respond to the increased cash-out prices. Indeed the short-term exchange price for EFA Block 5 on 10th December was only £51.06/MWh compared to £84.56/MWh and £73.68/MWh on the 9th and 11th December respectively. While we would attribute the limited response of the short-term energy markets in part to the unduly low level of cash-out, we would expect several features of the current market to limit efficient interaction between cash-out prices and forward energy markets. In particular, NGC's procurement of operating reserves in timescales different to those over which the energy market clears potentially distorts short-term energy market prices. We discuss this further in the following sub-section before considering some other potential constraints on efficient interaction between cash-out and forward markets in section 3.4.2 below.

3.4.1 NGC's Balancing Actions Are Not Adequately Reflected in Short-term Energy Market Prices

Energy and reserve markets are intrinsically linked. The planned level and time profile of energy production affects the availability of all forms of reserve and vice versa. In particular, the loading of individual units has a direct impact on the levels of regulating reserve available. In the opposite direction, NGC's procurement of reserve via PGBTs can also influence the amount of energy provided to the system, eg, if NGC buys the energy associated with bringing a plant to minimum stable generation to provide regulating reserve, or if the presence of a standing reserve contract precludes a plant's direct participation in short-term energy markets. Ideally, energy and reserve markets should be cleared over the same timescales to ensure optimal pricing of both energy and reserve as "joint products" procured from the same set of production plants. In particular, NGC's requirement for regulating reserve should effectively "remove" supply from the price stack available to short-term energy markets, to ensure the optimal allocation of generating capacity between energy production and reserve provision. Under the current system, however, reserve and energy markets clear over different timescales. Short-term energy markets effectively clear before gate closure with the majority (circa 95 per cent) of short-term trading and unit commitment decisions being made at the day-ahead stage. Although NGC only knows the actual level of regulating reserve available and the prices attached to that reserve (as implied by generators bid and offer prices) at gate closure, they have a good initial idea of the likely dispatch at 11am at the day-ahead stage. However, NGC typically only acts to secure operating reserve within day via the BM and PGBT actions.

This inconsistency in the clearing periods between energy and reserve markets will at times lead to significant inefficiencies and distorted cash-out and short-term market prices. NGC has financial incentives and regulatory obligations to purchase balancing services efficiently and economically and there is every reason to believe that within this framework NGC perform their role efficiently. However, minimising NGC's *purchase* cost for balancing services does not necessarily minimise the cost of energy production and reserve provision across the interlinked markets. With perfect information and foresight – for both NGC and generators - NGC incentives would be sufficient to ensure efficient trade offs. With perfect information, generators would also be able to assess the relative likelihood of being called and the likely market clearing prices for providing balancing services versus the prices available in short-term energy markets and make optimal decisions on whether to commit to provide energy or to wait and offer balancing services to NGC. However, in practice, the lack of perfect information on likely system developments means that even if NGC procures efficiently, generators are not offering prices and availabilities which reflect efficient trade offs between forward and balancing services markets. In particular, generators cannot realistically predict their likely output and the clearing price for offering balancing services. This results in generators preferring to arrange their planned output at the day-ahead stage, rather than optimising their output across day-ahead markets, balancing services markets and the balancing mechanism. (Operational drivers may also result in generators preferring to lock in their operation day-ahead rather than trade across the shorter-term markets.)

In a long system, the timing problem will not cause too many problems because significant regulating reserve is likely to be available anyway by backing down over-committed plant to part load. In a short system, however, the timing issue could result in serious problems. In particular, at times of emerging tightness in the market, short-term energy markets will clear below efficient price levels, because the associated erosion in the expected availability of regulating reserves has not been fully taken into account in determining forward prices (primarily because NGC does not purchase their reserve requirement at the same time as the energy market clears). The result will be inefficiently low levels of short-term energy prices and limited availability (and consequently higher prices) for subsequently procuring regulating reserve. At times when the market is very tight, this situation could mean that there is insufficient capacity synchronised to meet operating reserve requirements and demand disconnections become a possibility. These circumstances therefore allow cash-out prices to rise to very high levels, but without an efficient feedback of those prices to short-term energy markets. (In turn, this will provide inadequate longer-term price signals to ensure adequate levels of installed capacity.)

In addition, several other factors suggest that while NGC purchases efficiently their actions may not result in the wider efficient arbitrage required between the provision of balancing services and forward energy markets, if balancing signals are to be properly reflected in short-term energy prices. For example:

- Many of the actions designed to procure additional regulating reserve are undertaken through "off-market" bilateral PGBT's rather than on short-term exchanges with the result that these purchases fail to feed back effectively into short-term market prices. The lack of transparency to the wider market on why individual PGBT's are made exacerbates this problem.
- NGC may be reticent to purchase additional energy in short-term markets due to concerns about possible feedback to the pattern of dispatch, thereby creating instability in their expected pattern of reserve holding (and possibly undoing PGBT transactions).[6]
- Uncertainty about the likely availability of regulating reserve may result in NGC's failing to trade out fully the expected system imbalance. For example, NGC may make a PGBT sale to take a unit off to ensure adequate downward regulation on the remaining units, but without making the corresponding purchase in short-term energy markets. Conversely, NGC may synchronise a unit to provide regulating reserve and buy the minimum-stable generation under a PGBT, without making the corresponding sale in short-term markets. While this may be efficient procurement on NGC's part, the failure to trade out the associated energy imbalances means that short-term energy prices do not adequately reflect these actions.

These problems suggest that at times there will not be an efficient linkage between short-term energy markets and the markets for balancing services. In particular, at times of shortage, short-term energy prices often fail to respond effectively even when balancing prices rise to very high levels. In turn, longer-term markets will fail to capture the true value of generation at times of shortage.

3.4.2 Additional Constraints on Efficient Optimisation of Balancing Mechanism and Forward Market Operations

As noted above, NGC's procurement of reserves and balancing services is a source of particular concern about the interaction between balancing mechanism and forward markets. However, more widely, while dual cash-out pricing provides incentives for market participants to balance their own positions before gate closure it does so by discouraging market participants

[6] NGC made this point in its presentation on the PGBT process review at the Operational Forum on 5 March 2003.

from taking open positions into cash-out. This limits the extent to which market participants can make optimal choices between trading in forward markets, participating in the balancing mechanism or taking imbalances to cash-out. Even if balancing mechanism and cash-out prices rise to high levels, there is therefore no guarantee that forward prices will rise in response.

The lack of an efficient link between BM, cash-out and forward markets is borne out by experience in the market, where forward prices show remarkable inertia in the face of changing cash-out prices. On 10 December 2002, even with the emerging shortage and high expected cash-out prices, forward prices on the power exchanges were actually *lower* than the 9th and 11th December. Since the implementation of P78, the continuing divergence between SSP and SBP also provides *prima facie* evidence of a sub-optimal link between the BM and wider market. (Short-term market prices – and hence the reverse price - *should* be a direct function of the expected main price, but in practice this relationship fails to hold.) Against these concerns, we would note that:

- Dual price cash-out and pay-as-bid acceptances in the balancing mechanism have worked well. They have facilitated NGC's management of the system by preventing market participants from short-term speculation on cash-out prices and have limited the scope for the manipulation of cash-out, and consequently forward, market prices.
- Market participants with physical flexibility (predominantly generators, but also some demand sites) can choose whether to trade forward or bid into the balancing mechanism and should, in theory, therefore be able to arbitrage out any persistent differences between balancing mechanism prices and forward prices.

The problem therefore appears to be less with pay-as-bid acceptances and dual cash-out in itself, but more with the failure of physical market participants to optimise their deliveries between balancing mechanism and forward markets. However, it appears that the root causes may lie in the deficiencies of the balancing and cash-out regime highlighted above. In particular, we would expect market responses to cash-out price signals to be far greater in the event that cash-out prices were set efficiently and rose to the much higher levels required in the event of a serious shortage of generation. At the present time, therefore, we are undecided whether dual cash-out is itself a source of major *additional* concern *over and above* the serious concerns highlighted above in relation to failure of the main cash-out price to reflect marginal costs and the imperfect link between cash-out and forward markets.

3.5 Tagging Methodologies are Imperfect

Efficient cash-out prices which reflect the marginal cost and value of <u>energy</u> acceptances in the BM require efficient separation between system balancing actions and energy balancing actions. Ideally the effect of locational constraints, the procurement of operating reserves, ancillary services and dynamic constraints should therefore be stripped out of the calculation of energy cash-out prices, such that the cash-out price reflects the marginal cost or value of energy balancing alone. This can be a difficult task, particularly in a market without centralised scheduling and dispatch as with NETA. As noted above in the discussion of operating reserve procurement, many system-balancing actions typically have associated energy balancing consequences (and vice versa).

The current methods for separating energy and system actions work reasonably well most of the time. The de minimis rule and the Continuous Acceptance Duration Limit (CADL) tagging rules remove the impact of short-term actions more likely to relate to the profile of production within a half-hour rather than the half-hourly balance for the basic energy commodity (ie, MWh delivered flat over the half-hour). Constraint tagging – based on the Net Imbalance Volume (NIV) – also largely removes the impact of locational constraints on cash-out prices. However, the system of tagging is less robust to times when the system is short of generation capacity. At these times, the main price will typically be SBP based on accepted offers. However, the need to maintain operating reserves – even at times of capacity shortage – means that NGC will also accept bids in the BM at these times. The net result is that even if all available offers are subsequently accepted – as they had been on 10 December 2002 – the most expensive of those offers will be tagged out, resulting in cash-out prices which underestimate the true average cost of the shortage. Thus not only will weighted average SBP tend to understate the cost of shortage, but that average itself is likely to be lower than it would otherwise be.

4. Improving the Operation of the England and Wales Electricity Market

Section 3 above has highlighted several features of the current NETA arrangements which undermine the determination of cash-out prices and efficient interactions between BM, cash-out prices and forward markets. These features of the current NETA arrangements do not adequately meet the key electricity market design criteria identified in section 2.2 and the NETA arrangements as they stand are therefore unlikely to deliver appropriate market signals on the value of generation capacity at times of system shortage. The NETA design has profound implications for the level of generation security in the England and Wales electricity market and this is borne out by emerging evidence. The Drax contract in November 2002 and the problems on 10 December 2002 highlight problems experienced last winter. NGC is also forecasting a serious deficit in the Operational Planning Margin Requirement (OPMR) for next winter. There is therefore an urgent need to develop the NETA arrangements in the next few months to respond to these likely security problems.

Designing an electricity market to provide adequate security of supply is a difficult problem, involving a complex set of decisions on the design of the energy market and associated procurement of balancing services. There are several market designs that can achieve an optimal level of investment and, hence, the optimal level of security of supply and many trade-offs between short-term market efficiency and longer-term security concerns are involved. In short, there is no unique correct solution and any proposed solution is firmly in the economic territory of a "second-best" solution. The current NETA market design also constrains the choice of options if, as we assume, a fundamental overhaul to the current market arrangements is to be avoided. Nevertheless, several pragmatic and relatively simple steps would significantly improve the price signals provided within the current NETA framework and we set these changes out in further detail in the following sections.

4.1 Cash-Out Prices Should be Based on Marginal Costs

Weighted-average pricing and the current methods for tagging mean that cash-out prices do not reflect the marginal cost and value of electricity at times of shortage. The solution is to base the calculation of SBP (when the system is short) and SSP (when the system is long) on the marginal accepted <u>energy</u> offer. This would not require any significant change to the BM. In particular, the BM could continue to operate on a pay-as-bid basis and the current P78 mechanism of a main price defined by system length and a reverse price would be retained. This retains the competitive benefits of pay-as-bid pricing, while ensuring that cash-out prices continue to send appropriate signals to the forward markets.

The main challenge is to define an appropriate marginal cost pricing rule and to tag out actions that are accepted for system, rather than energy, reasons in a straightforward and transparent manner. This is particularly important with a marginal approach to setting cash-out prices, since a system action which is not correctly tagged out directly sets the cash-out price, rather than just contributing to a weighted average of all acceptances. The aim is to tag out:

- Acceptances driven by locational constraints;
- Acceptances required because of the dynamic profile *within* a half-hour rather than to meet an energy balance *across* the half-hour;
- Non-marginal, forced acceptances to meet dynamic constraints, eg, the acceptance of generating in neighbouring settlement periods to satisfy minimum on or minimum off constraints.

Given that most acceptances have both system and energy consequences, it is unlikely that <u>any</u> tagging methodology will give a uniquely correct answer in all possible circumstances. Nevertheless, even an imperfect approach to tagging with marginal pricing is likely to be significantly more efficient than the current system of weighted average pricing and imperfect tagging. Several methods for tagging can be considered. For example, an ex post unconstrained schedule of acceptances would yield a proxy for the marginal energy price by not including acceptances required by system or other operational constraints. However, such an approach would raise significant implementation difficulties in accounting for dynamic constraints – such as minimum on and minimum off times – which link acceptances across several settlement periods. This would require rules to look back (or look forward) to determine when units are effectively non-marginal because the dynamic constraints effectively result in "forced" acceptances in settlement periods outside of those for which the unit's output is committed. This would require any schedule to extend beyond the immediate half-hour, which may raise problems on prompt price reporting.

We therefore propose that the calculation of the marginal cash-out price should be based on actual BM and balancing service acceptances. SBP/SSP would be calculated as the marginal actual acceptance derived from a price stack made up of both acceptances in the BM and acceptances made before gate closure. (That is, there would be a "consolidated" stack made up of all balancing acceptances whether or not those were made before gate closure.) The current rules for tagging would be retained to remove the effect of system and other non-marginal acceptances, ie:

- NGC would continue to flag PGBT actions taken for system reasons and these would be excluded from the price stack;
- The current BSC rules relating to CADL, Arbitrage and De Minimis tagging would apply to the consolidated price stack to tag out short-duration, small volume and arbitrage actions.

The marginal price would then be calculated by applying the Net Imbalance Volume (NIV) to the residual "consolidated" stack of untagged offers and bids. This would require relatively little change to the BSC rules themselves since the existing tagging algorithms would remain largely unchanged. However, to ensure a consistent approach to acceptances within and outside the BM and to capture the full cost of balancing services in cash-out prices, the following three expansions to the current methodology are required:

- The tagging methodology should be extended to remove "undo" offer (or bid) acceptances on the same BMU. This is to address an anomaly with the current form of NIV tagging which would result in inappropriate tagging out of the marginal acceptance at times of shortage.
- NGC will need to report pre-gate closure acceptances on a disaggregated basis via BSAD to feed into the "consolidated" stack and the extended tagging methodology would also need to account for offsetting forward energy trades.
- To address the problems associated with standing reserve highlighted in section 3.3 above, we also propose that the consolidated stack should include an estimate of the full marginal cost of standing reserve acceptances which accounts for the option fees.

The following three sections explain our proposals in these areas in more detail.

4.1.1 The Tagging Methodology Should be Amended to Tag Out "Undo" Acceptances

The current method of NIV tagging has an anomaly where NGC may have accepted a bid on a unit (say a pre-gate closure bid to back down to provide operating reserve) but subsequently accepts an undo offer and subsequent offers from the same and other units because of increases in demand or shortfalls on generation. Under the current system the bid acceptance could result in the highest price offers from the offer stack being NIV tagged even when NGC has "used" all operating reserves and has accepted subsequent offers to meet an unexpected shortfall in generation. This could lead to serious understatement of the marginal cost in the event that NGC has used all available operating reserves and "blown through" the offer stack in the event of generation shortages.

The solution to this anomaly is to amend the tagging methodology to remove offsetting offer and bid acceptances on the same BMU from the bid and offer stacks <u>before</u> the application of NIV tagging. For these purposes, there would be no distinction between pre-gate closure and post-gate closure acceptances on the same BMU.

Netting on a BMU basis prior to the application of NIV tagging would continue to remove the highest price offers in the case of genuine locational constraints (where the offers and bids will be accepted on different BMUs in different locations) and where bids and offers have been accepted on different units to create operating reserves or for downward regulation. However, it will better reflect the underlying economics at times of shortage, since the "undo" offers are likely to be towards the bottom of the price stack. (Given NGC's obligation to balance efficiently, they are less likely to accept units with unfavourable undo prices to create operating margin.)

4.1.2 Reflecting Marginal Costs of Non-BM Actions

A marginal approach to price setting should treat pre-gate closure balancing actions on the same basis as BM actions. This will require associated changes in the BSAD methodology to ensure that the volumes and associated prices of PGBTs and forward energy trades (including system-to-system services) feed through into the cash-out price calculations on consistent basis with BM acceptances. We would see this working as follows:

- NGC would need to amend their trade capture and reporting systems to ensure that the prices and volumes associated with PGBT energy trades (and other pre-gate closure energy actions) feed into the price stack on a disaggregated BMU-by-BMU or trade-by-trade basis.
- NGC would continue to flag PGBT trades for system reasons to ensure that they are not included in the determination of energy cash-out prices.
- Offsetting BM and PGBT bid/offer acceptances on the same BMU would be removed from the bid and offer stacks in line with the amended approach to tagging described in section 4.1.1 above.

The tagging of "undo" acceptances out of the bid and offer stacks would also need to be extended to forward energy trades to ensure that the correct marginal price is calculated. For example, suppose that NGC takes the following actions:

Forward sale	100 MWh at £16/MWh
Forward purchases	100 MWh at £18/MWh
BM offer acceptance	100 MWh at £20/MWh

This could reflect a situation where NGC initially expects the system to be 100 MWh long, but subsequently buys 100 MWh forward as the system flips to being balanced and a further 100 MWh to counteract an eventual 100 MWh shortfall. In this case, the first 100 MWh purchase at £18/MWh would therefore "undo" the previous sale. The correct marginal price in this situation is £20/MWh. However, if all trades were fed through into the price stack – and these were the only trades – NIV tagging would remove the £20/MWh purchase. As with "undo" acceptances on BMUs this would therefore send a distorted signal that the marginal cost of meeting the system shortfall was £18/MWh rather than £20/MWh. This is clearly anomalous – a forward sale (effectively at NBP) when the system ends up short can have been neither for locational reasons nor for the creation of operating reserve. It is therefore inappropriate that this volume contributes to the removal of higher price purchases (whether forward or in the BM) via the application of the NIV tagging process.

The solution to this problem would be to remove any <u>offsetting non-locational forward energy sales</u> volumes from the reverse stack and to remove an equal volume from the bottom of the purchase stack. (Similarly if the system is long, any offsetting non-locational energy purchase volumes would be removed from the reverse stack and an equal volume from the top of sales stack.) In the example, the forward purchase at £18/MWh would be removed from the offer stack and the forward sale at £16/MWh would therefore be removed from the bid stack. There would then be no residual bid volume for NIV tagging and the price would be set at the marginal accepted offer, ie, £20/MWh.

4.1.3 The Opportunity Cost of Standing Reserves Should be Included in Cash-Out Prices

Section 3.3 highlighted the problems associated with smoothing the capacity fees to standing reserve – and other firm regulating reserves - over the periods during which those reserves are made available. While this may be an appropriate way of remunerating standing reserve, it smoothes cash-out prices unduly at times of shortage. The way to correct this problem is to calculate the likely underlying "spot price" that would have obtained had standing reserve not been procured via annual option contracts. Put another way, the spot price should reflect the "opportunity cost" of using standing reserve on those occasions that it actually generates.

Given that standing reserves are procured annually, the "opportunity cost" can be calculated as the utilisation fee plus the annual capacity fee averaged over those periods in which the standing reserve is expected to be called. For example, a generator might offer an utilisation fee of £200/MWh and capacity fee of £10,000/MW into the standing reserve tender. If NGC expected to use this unit for 5 hours per year, this would translate into an opportunity cost of £2,200/MWh (ie, £10,000/MW ÷ 5 hours + £200/MWh) on each occasion that NGC expects to use standing reserve. The current method of allocating the capacity fees, however, would only result in a Buy Price Adjustment of roughly £2.50 for just fewer than 4000 hours per year and the £200/MWh utilisation fee being reflected in SBP when the unit was running. This results in unduly low cash-out prices on the relatively infrequent occasions that standing reserve is actually used.

NGC refers to the expected opportunity cost as the "effective cost" in their summaries of the tender assessment process. NGC does not procure any standing reserve with an effective cost above a Maximum Contract Price (MCP), which is set around £3000/MWh. This underwrites our contention that the effective cost is the correct interpretation of the opportunity cost since NGC effectively assesses standing reserves against a "dummy source of infinite reserve having a zero availability fee and a utilisation price of MCP".[7] The MCP is therefore the opportunity cost for accepting standing reserve tenders and is a proxy for the "spot price" that would obtain were standing reserves not procured.[8]

We therefore propose that the effective cost of standing reserve calculated by NGC during the tender process should feed through into the cash-out price stack via the BSAD at those times that the standing reserve is actually used. This could be achieved by using a BMU-specific Buy Price Adjustment (BPA_{ij}), equal to the effective cost, would feed through into the cash-out price calculation via the BSAD data. When standing reserve BMUs are used, the unit-specific BPA would then be *substituted* for their utilisation price (ie, their BM offer price) in the price stack used to calculate SBP.

Payments to standing reserves would remain as now, based on the utilisation fee and capacity fee. However, cash-out prices would reflect the full, expected cost of using reserves on those occasions when it was actually called. NGC should adopt a similar approach for any firm regulating reserve procured and to standing reserves procured from non-BM sources. While NGC currently recovers the costs of non-BM standing reserve via BSUOS, the calculation of cash-out prices and BSAD does not currently include these costs. NGC should therefore amend the BSAD methodology to ensure that the effective cost of procuring and using these reserves feeds through into cash-out prices in the manner described above when the reserve is called upon. This would be achieved by feeding the effective cost and associated volume of non-BMU standing reserves via BSAD into the price stack used to calculate SBP. This would ensure that there was no distinction between non-BM and BM sources of standing reserve in the calculation of cash-out prices.

---

7 National Grid Standing Reserve Market Report for Contracts Effective from 1 April 2002 to 1 April 2003.

8 We understand that NGC has used a slightly different approach in the tender round for contracts from April 2003 that includes an assessment of the effective cost against expected levels of balancing mechanism acceptances. Although this changes the calculation of the maximum contract price, it does not affect our contention that the effective cost is the appropriate opportunity cost to feed into cash-out price calculations.

To aid market transparency, it will also be essential to ensure that market participants can factor the expected level of prices when standing reserves are used into their forward price calculations. NGC should therefore also publish information on the range of effective costs with associated tranches of capacity in the form of a standing reserve "look-up table" following each standing reserve tender. This would enable market participants to predict the likely range of cash-out prices for different levels of standing reserve utilisation.

4.2 NGC Should Procure Operating Reserves More Transparently

NGC's recent consultation on improvements to the PGBT process signalled a willingness to improve the transparency of that process. This should lead to increased efficiency in the procurement of PGBTs and a limited increase in transparency on NGC's management of reserves. However, as described in section 3.4.1 above, the interrelationships between energy and associated markets (including reserve) mean that reserve and energy markets should ideally clear in consistent timescales. To facilitate this, we propose that NGC should procure its expected operating reserve requirement in advance of gate closure. In addition to the tenders for fast and standing reserve, NGC would explicitly procure contingency reserve, regulating reserve and downward regulation via an open transparent screen-based method similar to that being considered as a development to the PGBT process. This would improve interactions between short-term energy markets and reserve procurement. There are many possible designs for this reserve market, but as a starting point, we would propose the following:

- NGC organise a day-ahead reserve tender for contingency, regulating reserve and downward regulation.
- Procurement would take place within a screen-based, transparent mechanism similar to that envisaged for PGBT transactions
- NGC organise a within day mechanism for changes to the day-ahead pattern of reserve holdings. (Ideally, this would be based on the same screen-based system used for the day-ahead procurement.)
- Participants would offer a capacity fee for providing reserve together with an associated utilisation price.
- The BSAD methodology would be used in the manner described above for standing reserves to compute an "effective cost" of using the reserves that would feed through into cash-out prices when energy is taken from those reserves.

4.3 Ofgem Should Clarify the Reliability Policy

Section 3.1 highlighted some gaps in the current reliability policy and, in particular, the absence of a proxy for demand side response in the event that demand has to be disconnected and the lack of clarity on the interpretation of NGC's rights and obligations to balance the system.

The reflection of the opportunity cost of standing reserve in cash-out prices would partially fill this gap with respect to demand side response. The reflection of the effective cost of standing reserve in cash-out prices would signal emerging shortages via cash-out prices when standing reserve is used. Prices would rise gradually above marginal production costs as generation gets tighter and increasing amounts of standing reserve are used. In the limit, when demand is actually disconnected prices should be expected to rise to (at least) the MCP (the ceiling price for NGC's acceptances of standing reserve). This would therefore amount to a version of operating reserve pricing where the costs of capacity are increasingly reflected in market prices when generating capacity gets short.

However, this change would not be sufficient to clarify the other identified gaps in the reliability policy including the interpretation of the obligations on suppliers and NGC and the theoretical and practical gaps in suppliers' obligations to serve their consumers. We would therefore urge Ofgem:

- to issue a guidance note on its interpretation of the current licence obligations on NGC and suppliers; and
- to initiate a consultation on potential changes to clarify and strengthen the regulatory framework to address the gaps in the reliability policy.