Appendix 5.1: Wholesale electricity market rules

Contents

Introduction

1. This appendix examines three aspects of the design of the GB wholesale electricity markets:

   (a) mechanisms governing the dispatch of wholesale electricity (centralised dispatch versus self-dispatch);

   (b) the influence on suppliers' and generators' short-run costs of the cash-out rules[1] as recently reformed by Ofgem as part of its electricity balancing significant code review (EBSCR); and

   (c) the interaction between the investment incentives provided by two separate regulatory mechanisms designed to ensure that sufficient capacity is available at time of system stress: the capacity auctions and the pricing of cash-out.

   We assess the degree to which these three aspects of the design of the GB wholesale electricity market might preclude a well-functioning market.

Self-dispatch versus centralised dispatch

2. The current dispatch mechanism in force in Great Britain, introduced by the NETA/BETTA reforms,[2] was designed as a self-dispatch wholesale electricity market. This contrasts with the system that it replaced, 'the pool',[3] which was

[1] If a market participant generates or consumes more or less electricity than they have contracted for, they are exposed to an imbalance price, or 'cash-out', for the difference. The two terms are used interchangeably in this appendix.
[2] See Appendix 2.1: Legal and regulatory framework.
[3] See Appendix 2.1: Legal and regulatory framework.

centrally dispatched. This section considers the impact of each design on competition.

3. In a centralised dispatch system, generators and flexible demand[4] tell the system operator (SO) the prices at which they are willing to supply to the system and the prices at which they are willing to reduce consumption. These bids come with detailed technical information of constraints in plant operation. The SO determines what it believes to be the least cost way of matching supply and demand and communicates a planned running order to each participant. Sometimes the operating instructions will be determined up to 24 hours ahead of production; sometimes it will be as little as five minutes before production. In determining the running order of plant, the SO also determines the system price in each period that is consistent with that running order. Centralised dispatch exists in the Australian national electricity market (NEM)[5] and in some form in most deregulated markets in the USA.

4. Despite central dispatch and the establishment of a spot price defined by market rules in all existing systems, most electricity is in actual fact bought and sold in futures and forward[6] markets where supply and demand determine prices for electricity at different terms.

5. Under a self-dispatch system, buyers and sellers of electricity contract ahead of time for their anticipated demand at prices that are bilaterally negotiated or determined through demand and supply matching on public exchanges. Generators and suppliers prepare operating plans for their anticipated physical behaviour or that of their customers. The parties communicate their anticipated physical behaviour and their contractual position to the SO.

6. The SO takes central control of balancing supply and demand close to real time, at a point known as ‘gate closure’. It is an intrinsic feature of modern electricity systems that at some point the matching of physical supply and demand is a natural monopoly activity that requires central control over operating decisions. This applies to self-dispatch systems too. In this sense, the GB system is not truly self-dispatched – there is always some point at which central control is asserted. After the fact, discrepancies between what

[4] ‘Flexible demand’ refers to consumers who have the flexibility to reduce consumption at short notice in response to market signals.
[5] In the Australian NEM, dispatch is determined 5 minutes ahead of time, rather than one day.
[6] The distinction between ‘futures’ and ‘forward’ markets in the context of the GB electricity market is that a forward trade is a trade for physical delivery, while a future trade is a financial contract written against a reference price. Forward trading is more prevalent in GB, while futures trading is more prevalent in the Australian NEM.

parties physically did (actual delivery or offtake) and their contractual positions are 'cashed-out' at prices determined administratively by the SO.[7]

7. In Great Britain, the SO[8] receives notification of the physical and contractual position of each party one hour prior to operation. It uses this information as well as its own forecasts to assess whether the system is at risk of imbalance. The SO will intervene if it predicts a discrepancy between the amount of electricity produced and demanded during a certain settlement period. The SO has the obligation to balance the system at minimum cost and has wide latitude in determining when and what it purchases. It also requires all parties who have notified the system that they will operate to also announce the physical constraints and financial parameters under which their plans can be altered by the SO. This is one element that allows the SO to determine a least-cost course of action in its balancing duties.

8. In order to assess whether self-dispatch systems impede effective competition, we consider in turn the main arguments that may inform our assessment:

   (a) self-dispatch reduces technical efficiency;

   (b) self-dispatch reduces price transparency; and

   (c) self-dispatch increases transaction costs for new entrants and smaller players.

Self-dispatch reduces technical efficiency

9. In most commodity markets, prices consistent with technical efficiency[9] are discovered through bilateral competition or, sometimes, through bids and offers on exchanges: a less efficient provider cannot profitably offer a price that is attractive to buyers if the more efficient firms are competing on price for sales. In the case of homogeneous commodities bought in wholesale markets by sophisticated agents, this process is likely to work quite well.

10. Traditionally in centralised, often monopolised and public electricity systems, technical efficiency was achieved centrally without the use of market mechanisms by calculation of cost-minimising operational plans. Some deregulated electricity markets retained the centralised calculation of optimal

---

[7] Effectively, any physical shortfalls or excesses compared to contract are 'made up', or balanced, by contracts with the SO in the process known as cash-out, ensuring that all electricity physically produced or consumed is also sold or bought. After gate-closure, the only counterparty to these trades in the GB system is the SO.
[8] The exact definition and duties of an SO vary from system to system. In the GB system, National Grid Electricity Transmission plc. carries out the SO role.
[9] 'Technical efficiency' refers to the property of minimum-cost production for the economy as a whole.

operation but required firms to compete by offering attractive input values to the centralised calculation (centralised dispatch). All electricity systems maintain some degree of centralised dispatch decision.

11. The evidence we have seen suggests that bilateral trading is leading to close to technically efficient operation of the system. Several parties have shared with us their modelling approaches based on cost minimisation by the SO and their close fit to actual prices. We have reviewed these models in the context of our work on unilateral upstream market power[10] and we find that their results are convincing. If bilateral contracting were leading to systematic technical inefficiency, we would expect to see this in systematic deviations of forecast and actual prices. We do not see these in the model calibration results. Our own wholesale price modelling[11] suggests that day-ahead prices are well forecast by a cost-minimising assumption.

12. InterGen has suggested that some of its combined-cycle gas turbine (CCGT) plant runs less frequently than less efficient plant owned by some of the large vertically integrated companies, and suggests that this is evidence of technically inefficient operation.

13. In order to investigate this claim, we asked Intergen for examples of specific periods in the year when some of its plant was not operating but when competitors it considered to be technically equivalent were operating.

14. Intergen offered an analysis of the following plants:

Table 1: Efficiency of Intergen CCGT plant and close substitutes

Plant name	*Owner*	*Year of operation started*	*Efficiency (%)*	*Technology*
Rocksavage	InterGen	1998	49.5	Alstom GT26 A/B
Coryton	InterGen	2001	49.2	Alstom GT26 A/B
Shoreham	Scottish Power	2000	49.5	Alstom GT26 B
Little Barford	RWE	1996	49.5	GE 9FA
Connah's Quay	E.ON	1996	49	GE 9FA
Rye House	Scottish Power	1993	48.5	Siemens V94.2
South Humber	Centrica	1999	49.5	Alstom GT 13 E2 with MXL upgrades

Source: Intergen. The quoted efficiency levels for competitors are estimates only, based upon the age and technology of the turbines.

15. Looking at the average load factors in peak periods for these plants in 2014, Intergen found 'that Little Barford (LBAR-1) and South Humber (SHBA-1 and SHBA-2), in particular, have significantly higher load factors, in the case of

[10] Appendix 4.1: Market power in generation.
[11] Appendix 4.1: Market power in generation.

Little Barford (RWE) nearly 70% compared to just over 20% for Coryton (COSO-1).'

And concluded that 'we would not expect the load factors amongst these plants to be so materially different over the course of a year.'

16. The evidence provided, however, does not entail that there is any inefficient dispatch. Based on InterGen's data, Coryton is less efficient than Little Barford and South Humber. Efficient dispatch would require a lower load factor for InterGen's plant, which is what InterGen observed. The evidence provided thus does not support the claim of inefficient dispatch.

17. The claim Intergen actually made about its evidence was a different one, about the sensitivity of load factor to efficiency. In a competitive portion of the merit order, it is entirely possible that small differences in cost will have drastic impacts on operations and therefore load factors.

18. Centrica agreed that Intergen's Rocksavage and Coryton utilisation appeared lower than one might expect from its reported efficiency alone. It pointed out that Rocksavage had suffered a number of prolonged outages in 2014. But Centrica argued that "there are many influences on a generator's price offers and plant dispatch decisions, apart from the relative efficiency of its plant."

19. SSE pointed out that it had also found observations similar to Intergen's in its own portfolio. It explained using , its own examples, the precise way in which average operating efficiency was not a good enough, detailed determinant of cost-minimising running patterns:

> Variable operation and maintenance costs and start-up costs can differ significantly by manufacturer and technology. Examples of this type of variation exist within different long term service agreements (LTSAs) for assets within SSE's generation portfolio: [✂] LTSA has low start costs and negligible variable operation and maintenance costs but high fixed charges; [✂] LTSA is based on equivalent operating hours for both start-up and operation and maintenance costs but with low fixed charges. When these factors are taken into account, the observed running of such assets reflects efficient dispatch.

20. Another possible explanation suggested by InterGen is that the vertically integrated firms are better at forecasting the likely costs of balancing that are spread over each half-hour and therefore better able to forecast the profitability of a plant operating or not in a given half hour. However, the forecasting of imbalance levels and of prices does not appear to offer any distinctive advantage to a vertically integrated operator, especially now that

near-term trading is sufficiently accessible and liquid. Moreover, the single-pricing reform to the cash-out price will further reduce any disadvantage from reliance on the system operator for imbalance purchases or sales.

21. The evidence and explanations provided by parties lead us to consider that Intergen has not identified outcomes that can be characterised as systematic departures from efficient operation.

22. We asked National Grid to consider possible sources of savings that might be seen from reverting to centralised dispatch. It concluded that there would not be substantial savings from the point of view of balancing the system. It also commented that in moving from the pool to NETA, it found generation asset owners were now more reluctant to switch plants off than National Grid had been as central dispatcher under the pool. National Grid hypothesised that plant owners may be able to factor in the additional maintenance costs implied by frequent starts and stops more accurately than could the SO under centralised dispatch rules, and that self-dispatch may in this sense be more technically efficient.[12] Centrica made a similar point, concluding that no centralised system of dispatch based on 'daily bids' from individual generators was ever likely to reflect the actual economic realities of each individual plant as well as a system of self-dispatch.

Self-dispatch reduces price transparency

23. The extreme case of a centralised dispatch system, like the Australian NEM mandatory gross pool, requires all supply-side and all demand-side parties to submit bids and offers. The market generates a market price that is based on all anticipated market activity and is publicly available. In other systems – like ERCOT and NordPool – bidding into the pool is not mandatory but produces cleared prices based on the bids and offers that are submitted.

24. One advantage claimed for centralised dispatch has therefore been the public availability of a market price based on all (or a substantial proportion) of physical trades. Part of the value of a mandatory market comes from the fact that everyone can be confident that the price is the result of supply and demand matching in the whole market. If the price is the result of a market whose functioning is regulated as a public good rather than for private convenience, the resultant price can more naturally provide a firm and trustworthy reference price for policy.

---

[12] See National Grid (January 2015), *Would it be more efficient/less costly for National Grid to manage all dispatching?*.

25. We have found that for most purposes prices are transparent in the GB wholesale electricity market.[13] The N2EX and APX exchanges publish day-ahead electricity auction prices. Approximately 40% of total electricity generation goes through these auctions. The provisional findings set out in Sections 4 and 6 suggest that parties do not have the ability or incentive to make this price systematically diverge from a competitive spot market price. This suggests that the price signal from these auctions is likely to be robust. The N2EX and APX bids and offers are already used for the regulated purpose of determining EU-wide day-ahead prices and allocating interconnection capacity across the EU. It is far from clear that mandating that all electricity be traded in the day-ahead market would improve the quality of the price signal that is generated by the N2EX and APX exchanges.

26. Prices of individual trades in the forward market are available for a modest fee from Trayport, a screen-based trading software provider that most traders use. After the day-ahead market has cleared, adjustments to contractual positions are typically made through Trayport in bilateral trades. The prices of these trades are available to participants and subscribers. Our analysis of trading data suggests that 3% of energy traded externally one day ahead of delivery or less is traded through private bilateral contracts that are not visible to all participants. We do not consider that this is a material degree of opacity.

27. Real-time imbalance prices are made public, as are the balancing mechanism (BM) bids that went to determine those prices. The reforms to imbalance prices that are anticipated in the next three years – and particularly the move to a single imbalance price (see paragraphs 57 to 60 below) – should ensure that the imbalance price in most periods is a good measure of a real-time spot market price. In this sense, there will be, post-reform, a market price based on the real-time, mandatory centralised matching of supply and demand that applies to the whole market. It is a price based on all bids and offers in the market in the sense that anyone who has notified to the SO that they will be producing or consuming electricity is mandated to offer financial and technical parameters for adjustments to those positions.

28. In relation to price transparency, the difference between the Australian NEM, characterised as a mandatory gross pool, and the GB system after the proposed EBSCR reforms, is very slight and rather technical in nature. Generators are mandated to bid into the GB balancing mechanism just as they are mandated to bid into the NEM in Australia. In both systems, these bids are used by the SO to build a supply curve and to generate a price used in almost real-time purchases and sales. In both the GB system and the NEM,

[13] See Appendix 6.1: Liquidity.

most trading occurs outside the real-time market in forward and futures markets, both in brokered bilateral trades and on exchanges. This occurs, as it does in many markets, because of parties' desire to add predictability to cash-flows. Future trading adds a layer of obscurity to purchase costs, but there is no difference in that respect between the Australian NEM and the GB system. In practice, the difference in terms of price formation between the exemplars of 'self-dispatch' and 'mandatory gross pools' tend to disappear.

29. For all these reasons, we do not believe that there would be a large advantage to competition from the point of view of increasing price transparency by reverting to centralised dispatch.

Self-dispatch increases transaction costs for new entrants and smaller players

30. A separate advantage claimed for a centralised dispatch system is that it provides a simple route to market for energy: a generator knows that it can sell its output by bidding into a pool; a supplier can buy energy from the gross pool. In the case of a mandatory pool, the entire market participates, so the depth of the market is maximised.

31. Under a self-dispatch system, parties are responsible for finding generators or suppliers with whom to trade. This requires, in-house or outsourced, teams of buyers and sellers and may be more complex than participating in a pool.

32. However, even in centralised dispatch systems with gross pools, most of the trading takes place in the forward markets that lead up to bidding in the gross pool. This arises from the corporate need for prudent risk management.[14] The relevant comparison to assess transaction costs should not, therefore, be between having no need for a trading team versus needing a full trading team, since both self- and centralised dispatch systems typically require participants to have trading teams.

33. Once again, the difference between centralised dispatch systems and self-dispatch systems starts to become less clear the closer we look at the details of each, as each system requires mechanisms that are characteristic of the other system. For instance, centralised dispatch systems require a procedure for dealing with unanticipated changes to production or demand – they need some sort of BM – while self-dispatch systems need some form of central control in real time (ie it becomes a centralised system close to real time). The nub of the difference is the number of opportunities that generators and suppliers have to contract with the SO for energy. In the GB system, this

[14] See Appendix 6.1: Liquidity.

happens in the BM after gate closure; in centrally dispatched systems, this also happens then, but in some cases also before that, perhaps a day before.[15]

34. Participation in spot markets in Great Britain involves low transaction costs. The APX and N2EX auctions allow day-ahead trading on a very similar basis to that which would be provided by a gross pool.[16] Moreover, the reforms to the imbalance price regime (especially the elimination of 'dual pricing'; see paragraphs 57 to 60) mean that reliance on the centrally cleared BM for energy will no longer be unattractive by design. This will provide a further low transaction cost option for buying or selling electricity.

35. In the past, the regulator has been concerned that there might be insufficient incentives for parties to contract bilaterally and has therefore encouraged contracting by making imbalance commercially unattractive. That design choice could be characterised as increasing transaction costs through more bilateral contracting in order to reduce reliance on the natural monopoly system operator for balancing: the more self-balancing, the less the need for alternatives like Short Term Operating Reserve (STOR) to manage imbalances. The anticipated move to a single imbalance price will reduce the incentive to bilateral trading, although some elements of the reform (as discussed in Annex A) may reintroduce some elements of them.

Cash-out and the balancing mechanism

36. In this section, we describe and discuss the relationship between cash-out and balancing following the EBSCR in the GB wholesale electricity market. In particular, we examine how the reformed cash-out rules use competition to minimise short-run costs.

37. Where more electricity is generated than consumed, or vice versa, system frequency may rise or fall to a degree that requires central intervention (an imbalance). This can be the consequence of an unforeseen peak or an unexpected fall in supply or demand (eg due to weather conditions or technical failures in the system or by particular power plants). In order to prevent imbalances, the GB system of maximum self-dispatch becomes a centralised mechanism close to real time. This is required by the nature of the

[15] This is not true of the Australian NEM, which might be characterised either as operating a centrally dispatched system or as operating only a mandatory gross balancing mechanism with no gate closure.
[16] The N2EX requires parties to post collateral for their trades. This may be a substantial cost, but we have not found evidence that it is an undue cost. A day-ahead pool would also need to have some insurance mechanism against a party's inability to make good on its commitment.

coordination problem involved in maintaining an interconnected grid that is safe and stable at reasonable cost.

38. For this purpose, National Grid as SO calls on the following tools at its disposal:

(a) It forecasts supply and demand for every instant in the near future (an important point in this process is the determination of its reserve requirement of 4 hours of production; 4 hours is the time it takes many gas and coal plants to be switched on).

(b) One hour before production (at 'gate closure') it requires physical generation and offtake plans from all parties (Final Physical Notification), which is one input used in making the forecast of supply/demand balance.

(c) It develops a plan to call on any of its own supply and demand options, which include (from more to less common):

(i) accepting bids to increase or decrease output or demand in the BM in the hour before production (after gate closure);

(ii) buying supplies in the open market before gate closure; and

(iii) calling on previously contracted options for balancing services, like STOR.

39. Balancing has two aspects: the physical activities of National Grid and the financial settlement by National Grid's subsidiary, Elexon. The question of market rules in this area focuses on impact of the methodology of financial settlements on both the physical balancing activity and the wider market.

40. The costs of the physical balancing activities need to be apportioned. In a process that can be thought of as a monetary reflection of the physical process of balancing, parties submit Final Contractual Notifications to the SO[17] up to gate closure. These describe the contractual positions of parties: how much electricity each has bought or sold up to one hour ahead of a half hour balancing period. If there is found to be a deviation between physical production or consumption and quantities sold or bought before gate closure, the party pays or is paid for the electricity involved with the SO as counterparty.[18]

---

[17] The notification is actually to Elexon, the subsidiary of National Grid that settles accounts.
[18] The settlement process can be long and is made of a succession of approximations. It can take up to 14 months for accounts to be settled fully, and even then, in the absence of complete coverage by smart meters, payments may not reflect physical activity. See Appendix 8.6: Gas and electricity settlement and metering.

41. Generators and suppliers who are deemed to have contributed to an imbalance (eg through a failure to comply with their contracted delivery or offtake) are charged an imbalance price ('cash-out') if they are 'short' (produced too little or consumed too much relative to contract); and paid an imbalance price if they are 'long' (produced too much or consumed too little).

42. In any settlement period, a generator's revenues[19] from the production of electricity is equal to its revenues under contract plus or minus the cash-out price multiplied by the uncontracted quantity. If the generator has overproduced relative to contract, this last is a positive; if it has underproduced relative to contract, it is a negative. Similarly, a supplier's purchase costs for electricity are equal to costs under their contracts plus or minus the cash-out price multiplied by uncontracted quantity. If the supplier has consumed more than contracted, this last is positive; if it has consumed less than contracted, it is negative.

43. As one of several additional tools to maintain the capability of achieving an energy balance on the electricity transmission network (given the potential for various eventualities that will cause imbalance between approximately 4 hours ahead of time to real time), National Grid maintains STOR contracts whereby the counterparty undertakes to make available a contracted level of power when instructed by National Grid, usually with the requirement that this be available at very short notice (approximately 20 minutes).[20]

44. STOR contracts are procured via a competitive tender process with three tender rounds per year. National Grid pays an availability payment to STOR service providers, which is paid regardless of whether they are asked to produce, and a utilisation cost in case of actual delivery. STOR providers agree to make available capacity to National Grid and face contract penalties (depending on terms) if the capability cannot be made available: for example, because they are producing energy for other parties. In practice, this means that STOR capacity is reserved for use by National Grid, although the contracted parties retain commercial discretion to use the capacity elsewhere.

45. This discussion will lead in the next section to an examination of the combined impact of the Capacity Market and cash-out rules on longer-term investment

---

[19] This abstracts from other, ancillary, services that the generator can supply to the system operator and from balancing market revenues. A more complete description would be this: Generator Revenues = Contract Revenues + Balancing Service Revenues + Cash-out Price × Imbalance Volume, with Imbalance Volume = Energy Produced – (Energy Bilaterally Contracted + Energy Contracted in the Balancing Mechanism). The details are given in ELEXON (2014), *Imbalance pricing guidance*.

[20] The other tools are smaller in capacity and more limited in their uses. Details can be found at National Grid *What are reserve services?*.

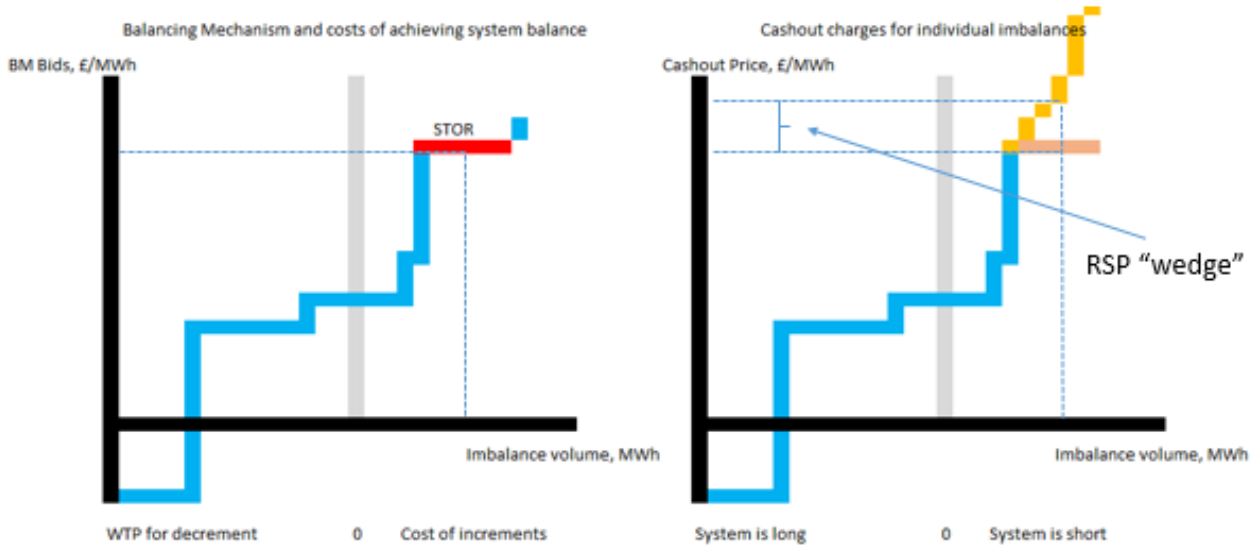
incentives. We will then consider the impact of the cash-out rules on 'balancing efficiency'.

An outline of balancing and cash-out

46. The Balancing Mechanism (BM) is the process by which National Grid procures and rewards the energy that it needs in real time (and up to 90 minutes before that) to balance supply and demand. The cash-out mechanism relates to the conceptually separate question of what parties are individually charged or paid for electricity that they use or produce without having a pre-gate closure contract to do so. We explore in this section how the pricing of cash-out and the likelihood of an imbalance influence the behaviour of market participants before and after gate closure, and how the mechanisms outlined above are using competition to minimise market participants' short-run costs.

47. The relationship between the pricing under BM and under the cash-out rules as reformed following the recent EBSCR (which is anticipated to come into force in a gradual way over the next three years) is shown schematically in Figure 1 below.[21]

Figure 1: Balancing mechanism and cash-out prices



Source: CMA analysis. WTP in the figure above means 'Willingness to Pay'.

[21] This diagram abstracts from many detailed elements of the relationship between balancing mechanism and cash-out, for example: STOR can be used at times outside periods of system stress; 'tagging' of actions; cost recovery and the Residual Cash-flow Reallocation Cash-flow (RCRC) 'beer fund'; and several additional sources of fast response. The diagrams also abstract from the auction design of the balancing mechanism, which, as a pay-as-bid auction, will not reveal balancing costs in the way assumed in the diagrams. These complications are not central to the arguments that follow.

48. The left-hand diagram shows the cost of achieving system balance in the context of the BM auction process. The very short-run supply curve for wholesale electricity is represented as the blue curve. The grey line represents the contracted demand curve (ie the expected point of system balance just before gate closure). The fine dotted lines show what the cash-out price would be if imbalance were such as to require the use of STOR.

49. On the assumption that parties are aiming to be balanced and that near-term markets are liquid, National Grid would not be expected at gate closure either to need to buy or sell energy (ie under these idealised assumptions, the grey line would represent both the net aggregate contracted position and the physical position of all market participants).[22]

50. However, there are always unexpected events on the supply and demand sides between gate closure and delivery which may cause an imbalance, requiring National Grid to buy or sell electricity through the BM. The extent of aggregate imbalance will determine the actions that need to be taken (and therefore the short-run marginal cost of National Grid's intervention to balance the system).

51. The short-run marginal cost of energy for balancing is given by the point of intersection of the actual demand for electricity (ie the out-turn imbalance) and the blue and red curve on the left-hand diagram. The blue portions of the curve represent actions that generators bid into the mechanism, and the red portion of the curve represents capacity available to National Grid under STOR contracts.

52. Individual parties may or may not themselves be in balance. For small parties, the probability of imbalance is independent of the overall system balance (while the imbalance of large parties is likely to cause overall system imbalance).[23] If a party has under-generated or over-consumed compared to its contracted volume, it will be charged for that shortfall of energy at 'system buy price'; if a party has over-generated or under-consumed compared to its contracted volume, it will have to sell that extra energy at 'system sell price'. These cash-out prices are derived largely from the weighted average prices of

[22] We relax some of these assumptions in some of what follows. Until now, it has been thought that parties prefer to over-contract and the expectation in most periods is therefore that the system will be long rather than balanced. We consider further below the degree to which this behaviour might change as a result of the proposed new rules for cash-out pricing.
[23] In what follows, this is the definition of a 'small' or 'large' party.

the offers and bids accepted by National Grid through the BM. This is shown in the right-hand diagram.[24]

53. Under the EBSCR rule changes, the single imbalance price will be set, as shown in the right-hand diagram, by the intersection of actual demand (ie overall system imbalance) and the supply curve.[25] This follows the supply curve for the BM over most of the range. However, when the system is short at high levels of demand and STOR comes to be used, the rules will introduce a wedge with the balancing market cost (in yellow in the diagram). This is known as reserve scarcity pricing (RSP) and is discussed at some length in paragraphs 104 to 121. The diagram shows an example in which RSP introduces a wedge between avoidable cost of generation and cash-out price.

54. In extreme cases where National Grid is not able to balance the system by increasing supply through the balancing auction and STOR contracts, under the proposed EBSCR reforms it will force some consumers to consume less energy (ie there will be blackouts or brownouts), and the imbalance price will be set administratively at £6,000/MWh.[26]

Description of the components of the EBSCR

55. The EBSCR reforms will modify cash-out pricing in four major aspects, each of which has been briefly touched on already and is described at greater length below.[27]

(a) A move to **single imbalance price**.

(b) A move to making the cash-out price in all periods equal to the cost of the 1 MWh most costly action in the BM (known as 'price average reference volume of 1 MWh', or PAR1), which is a narrowing of the base for the calculation from the previous 500 MWh (itself a narrowing from the original design, which was a simple average cost of all balancing actions).

---

[24] The cost of actions is not always reflected in cash-out prices and the SO goes through a complex 'tagging' procedure to determine which actions are properly energy imbalances rather than locational or other system-related effects. We abstract from these features of the mechanism in our analysis.

[25] This change will come into force by winter 2015/16.

[26] Strictly speaking, the exact procedure by which this occurs is slightly different. The EBSCR proposals do not include automatically setting the price to VoLL or RSP. Rather, one or more of the actions in the stack of actions used to calculate the cash-out price are to be repriced, at either VoLL or RSP, as appropriate. This stack of actions then undergoes a number of steps to remove certain actions – known as flagging and tagging – before price average reference (PAR) is applied and the price is determined. The impact of flagging and tagging – in particular of NIV tagging and SO constraint flagging – may mean that prices will not rise to VoLL even when there are blackouts or brownouts. See ELEXON (2014) *Imbalance pricing guidance*.

[27] Ofgem told us that 'EBSCR consists of an integrated package of several elements', the implication being that it should be assessed as a whole. In the sections that follow, we seek to assess each individual aspect of the package, noting interrelationships between aspects.

(c) A move to re-price STOR actions (typically periods of tight short-run margins due either to high demand or to supply disruptions)[28] to the probability of lost load (a measure of how stressed the system is, known as 'loss of load probability' (LOLP)) multiplied by £6,000/MWh (the putative 'value of lost load' (VoLL)),[29] if this is greater than their utilisation price. This is known as 'reserve scarcity pricing', or RSP.[30]

(d) A move to price disconnection or voltage reduction actions equal to the VoLL.[31]

56. The short-run incentive properties of each of these reforms are described in more detail below.

57. Single imbalance price is the proposed rule by which there is to be a single price for contractual imbalances. For example, if the system is short and a generator is producing more than contracted, it will receive the same price for its electricity as that paid by a supplier who has not contracted enough electricity. This rule is replacing the current dual imbalance price rule whereby actors who were long when the system was short or vice versa, (and were therefore contributing to the rebalancing of the system), were effectively penalised – or at least not rewarded – for doing so.[32]

58. The system of dual pricing was designed with the fear that parties might not have sufficient incentives to try to balance their supply and demand positions through bilateral contracts ahead of gate closure. The unattractive charge for beneficial imbalances was designed to encourage parties either to contract ahead of time or to participate in the BM but not to rely on cash-out as a market of last resort (ie taking a long or short physical position into the post-gate closure period voluntarily). The logic of doing this is one of 'balancing efficiency': of making the natural monopoly activity of centralised balancing as efficient as possible. This is discussed further in paragraphs 108 to 125 with

---

[28] Periods of tight margins are periods when STOR is likely to be used. However, STOR is also used outside of very tight periods. The SO has discretion to use a STOR plant over a balancing mechanism plant when it is more efficient to do so. STOR may even be used when the system is overall long. RSP, however, is likely to set cash-out prices only in periods when the margin is tight.

[29] LOLP measures the probability that the system will suffer an interruption. At times of high demand, the system loses resilience in that a power station breaking down could lead to an inability to find enough replacement capacity rapidly enough. This is the sort of situation when LOLP rises to being close to 1. LOLP is typically calculated by simulating the system. The VoLL represents the willingness to pay for an incremental MWh at times of system stress – it is the amount that the consumer of the last MWh is willing to pay to avoid being cut off. The value cannot be measured directly in any sense and is typically estimated once and for all using survey techniques.

[30] As noted above, the cash-out price calculation is applied to the stack of actions.

[31] There is a transitional period during which this will be set to £3,000/MWh and it will settle at £6,000/MWh

[32] This is achieved in the current system by those in 'helpful' imbalance being charged (if short) or paid (if long) an administrative price (the 'market index price') that was designed usually to be more (if short) or less (if long) than the corresponding payment or charge incurred in the balancing mechanism.

respect to the RSP reform. Suffice it to say here that, absent a concern over the efficiency of natural monopoly regulation, a single price in balancing appears to be uncontentiously good market design.

59. There is some evidence that the reform will be beneficial to smaller generators, to renewable producers and smaller suppliers who tend to be more reliant on cash-out than the large vertically integrated players. To a first approximation, we can consider that a small player aiming to be in balance will randomly find itself long or short with the same probability;[33] in the long term, under the single imbalance price, any losses made when contributing to the overall system imbalance should be offset by gains made when helping to solve it. Relying on cash-out as a market of last resort is no longer loss-making **by design**.[34]

60. Some small suppliers rely, even under current rules, to a much greater extent on cash-out than do the larger firms. This is plausibly because the transactions costs of being involved in the on-the-day bilateral markets are high. The move to a single price will make cash-out relatively more attractive for these parties.[35]

61. PAR1 is a rule change by which the calculation for the cash-out price outside times of system stress will be determined by the average cost of the last 1MWh of balancing actions taken. This will be introduced gradually, with 50 MWh being used next winter and 1 MWh introduced in 2018. This contrasts with the current rule by which the price is determined by the average of the last 500 MWh of actions taken (PAR500) PAR1 is described as making the imbalance price 'more marginal'.

62. RSP is a proposed rule change for cash-out prices in times of low short-term capacity margin, which will directly affect cash-out pricing when STOR contracts are called upon by National Grid. The SO typically seeks to contract under STOR ahead of time for between 2.2 and 2.3 GW of capacity.[36] As mentioned above, STOR providers are paid an availability payment by the SO and also, when called upon by National Grid to deliver electricity, a utilisation payment intended to cover its operating costs when it actually produces.

---

[33] A small player's own imbalance will not have a significant effect on system imbalance, hence the 'fair bet' involved in cash-out.
[34] Naturally, it still requires that these small players have sufficiently deep pockets or credit lines to balance out runs of bad luck without running out of liquidity.
[35] This is confirmed in Ofgem (2014), *Further analysis to support Ofgem's updated impact assessment*, Figure 3, which shows smaller suppliers benefiting from EBSCR.
[36] National Grid *STOR market information report: tender round 24*.

63. The RSP rule is expected to lead to a substantial net increase in cash-out prices at times of very tight short-term margin. Under the current system, there is a formula that averages out availability costs over periods when STOR was used historically. The cost of availability is thus reflected in the cash-out price, but not necessarily exactly when STOR was used. When the RSP rule bites, the cash-out price is raised administratively beyond the BM utilisation cost.[37]

64. VoLL is an administratively set price applied to cash-out payments if blackouts or brownouts occur for reasons of energy imbalance. This move to set prices to £6,000/MWh in those circumstances is a natural extension of the thinking behind RSP: if the LOLP is 100%, the cash-out price will be equal to the VoLL. [38]

65. Ofgem has argued that it was increasingly important to have the right incentives in system balancing because 'balancing costs incurred by the SO reached approximately £850m last year [2014] and are expected to rise substantially in future'. The overall EBSCR package is designed to improve balancing efficiency and to provide adequate signals for short- and long-run operational efficiency. The analysis presented in paragraphs 104 to 113 argues that the RSP and VoLL rule changes will interact with the Capacity Market (CM). We consider in paragraphs 114 to 121 the interaction between the CM, the cash-out rules and 'balancing efficiency'.

Use of competition to minimise market participants' short-run costs

Impact of PAR1 on short-run costs

66. We have considered three arguments concerning the move to PAR1 suggesting that the cash-out rules as reformed by ESBCR will not minimise short-run costs.

67. Stephen Littlechild has argued that PAR1 was not necessarily 'more marginal' because balancing actions are not necessarily simply incremental – they may be sequential. They may even be forward-looking and reflect expected imbalances in periods outside the period in which the action is taken. One solution to making balancing prices more clearly reflective of incremental energy costs is to reduce the settlement period from 30 minutes to something

[37] Subject to the flagging and tagging process mentioned above.
[38] Currently, demand disconnections in themselves have no impact on prices at all.

shorter.[39] The Australian NEM and ERCOT in Texas, for example, both use 5-minute intervals for imbalance price calculation. However, this design will make even worse the problem that actions taken in one period will be for purposes of balancing in another period.[40]

68. George Yarrow submitted that one of the original rationales for using an average price over a large number of actions was that this made the price less easy to manipulate; in the reformed cash-out pricing, with the cash-out price being calculated on the basis of actions amounting to 1MWh, it would be possible for a generator to learn that it tended to be a price setter in certain circumstances and might therefore be able to change its BM bids to take advantage of what would, in effect, be a lower price-elasticity of demand.

69. We have received arguments from Utilita, Ecotricty, Haven and First Utility that the move to PAR50 and subsequently to PAR1 would disadvantage smaller players who were more reliant than larger operators on energy purchases in cash-out.

70. In response to the first question, Ofgem has put forward two defences of PAR1:

(a) It carried out extensive qualitative analysis of the obstacles associated with a fully marginal price, including the issues raised in the working paper of accurately identifying the marginal cost action taken by the SO and the exercising of market power. We have examined Ofgem's qualitative analysis in its draft policy decision.[41] Ofgem argued 'tagging and flagging' actions, whereby certain bids were excluded from the calculation of cash-out, led to less sharp prices than would a simple PAR1 rule. While this shows that Ofgem has considered the problem, it does not, in our view, actually resolve the questions posed by Professor Littlechild. These can only be resolved by careful empirical work which has not been done.[42]

(b) Ofgem argued that the reform is being phased-in, with an opportunity to learn from the experience at PAR50; should this demonstrate that there are real problems with further tightening, the modification could be revisited. We suggest that full advantage of this phasing be taken and that Ofgem should use the opportunity of the move from PAR500 to PAR50 to

---

[39] We note that ELEXON has been asked by Ofgem to undertake initial analysis of the impacts, costs and benefits if the European Network Code on Electricity Balancing (EBNC) requires GB to adopt a 15-minute Imbalance Settlement Period (ISP).
[40] RWE points out the 5-minute settlement can also exacerbate market power problems by giving generators more opportunities to exploit momentary positions of dominance.
[41] Ofgem, (2013), EBSCR Draft Policy Decision Impact Assessment, Paras 4.13–4.15.
[42] RWE noted that a move to shorter settlement periods, as in the Australian NEM, was not necessarily a panacea and that this could exacerbate market power problems.

do a careful empirical analysis of the likely effects of a further move to PAR1.

71. On the point of the greater manipulability of a PAR1 price, National Grid has countered that any attempt to increase an offer price (or reduce a bid price) in the BM might result in the price of the action being removed from the energy imbalance price stack through 'tagging and flagging'. This would limit the extent that any individual could know that they would set the imbalance price. One way of interpreting this point by National Grid is to observe that it has the ability to learn about the manipulation and the discretion to counter it. With its responsibility to minimise the overall cost of balancing, it is not constrained in any mechanistic implementation of a least cost algorithm. Thus, the present system seems less prone to the sort of micro-manipulation of advantage than was the previous pool. We accept the view that this considerably reduces the increased risk of manipulation because of a move to PAR1.

72. We consider that the argument that the 'sharpening' of imbalance prices, of which PAR1 is one component, is a particular disadvantage to smaller players does have some merit. However, we noted that PAR1 in combination with a move to a single price may have a relatively small impact on smaller players because they can be expected, in the new regime and outside of RSP periods, to benefit approximately as frequently as they lose.

Impact of RSP on short-run costs

73. Ofgem has argued that RSP addressed defects in market efficiency that created the following symptoms:

    (a) Inefficient short-term trading and dispatch.

    (b) Dampened signals for interconnector imports during scarcity.

    (c) Inflexible capacity mix.

    (d) Dampened incentive for demand side response (DSR).

    Points *(b)* and *(d)* are about the price signals required to induce specific behavioural responses, while points *(a)* and *(c)* are what Ofgem calls 'balancing efficiency'. We consider each in turn.

Balancing efficiency

74. In Annex A to this appendix, we present a simple conceptual model that captures the essential mechanism by which RSP has effects like the ones

described by Ofgem. In essence, we can imagine the situation as being the following:

(a) The CM ensures that adequate capacity with a readiness to produce of 4 hours is available.

(b) The requirements of short-term flexibility are that some sunk investment expenditure is required above the cost of capacity to make plant responsive within 4 hours.[43]

(c) We simplify the market so that there are just two ways of recovering that sunk cost:

(i) A generator could sell the flexible capacity to the SO under a STOR or STOR-like contract.

(ii) A generator could contract privately to sell the output before gate closure at a negotiated price.

75. In our conceptual model we compare investment, total system cost, and the size of the SO (measured in terms of quantity of STOR purchases) in two different cases: the first, imbalances are expensive and parties prefer to pre-contract flexible capacity (what in the model we call the 'INVEST' scenario, which is the 'RSP' case); in the second, imbalances are priced at incremental cost and the SO purchases the required amounts of STOR (the scenario that we call 'SO' in the model).

76. The results of the model are intuitive:

(a) INVEST has a smaller role for the SO because more capacity is purchased bilaterally and STOR is therefore required less frequently.

(b) In the SO scenario, the SO invests in less capacity overall because it uses information closer to real-time on which to base its purchases; there are cases when INVEST contracts for capacity 'too early' in the sense that the predicted imbalance 4 hours out is resolved before actual production without the need for the contracted capacity.

(c) Ignoring the inefficiencies of natural monopoly regulation – and this is a crucial caveat – total system cost is smaller under SO than under INVEST because SO does not induce investment based on fear of balancing costs in excess of short-term incremental cost.

---

[43] We can imagine this as being either investment in reliability, in flexibility, in better forecasting or actually in an entirely new plant; the exact technical nature of the flexibility investment is not important.

77. This conceptual exercise suggests in our view that Ofgem's argument that RSP can improve balancing efficiency is correct: it can lead to a substitution between bilateral contracting and STOR purchases.

78. In order to assess the likely magnitude of this effect, we asked National Grid to describe the process it actually goes through in deciding how much STOR to purchase and how much to rely instead on bilateral contracting and BM bids. National Grid describe the following process:

(a) The quantity of STOR tendered for is fixed with consideration of two factors:

(i) National Grid assures itself that it has enough STOR capacity to cover the pre-defined **largest loss event** – currently set to be 1.8GW, corresponding to the loss of half of the anticipated Hinkley C capacity.[44]

(ii) National Grid forecasts physical availability on the system, anticipated balancing requirements and anticipated balancing market conditions; these forecasts allow National Grid to contract for STOR over and above 'largest loss' if it believes that this will be cheaper than procuring balancing energy in the near-term or balancing markets. This is a commercial decision by National Grid[45] which appears, within the confines of the natural monopoly regulation, to have strong incentives to minimise the cost of overall balancing. Currently, National Grid is buying 500MW of STOR over and above the 'largest loss' requirement.

(b) 24 hours before a production period, National Grid continually monitors supply and demand conditions and has a running forecast of its reserve requirement; there is a continually updated plan for how imbalances might be resolved and that plan sometimes entails purchases or sales in the market.

(c) Four hours before production, National Grid determines the system's required operating reserve for a production period; it forms a plan as to how much it can expect the market to resolve possible imbalances, how

[44] See Ofgem, (2011), *National Electricity Transmission System Security and Quality of Supply Standard (NETS SQSS): Review of Infeed Losses (GSR007 as revised by GSR007-1)*.
[45] As part of its natural monopoly regulation, National Grid agrees with Ofgem an anticipated balancing cost for two years; National Grid can keep a proportion of savings on that cost; if the savings are due to improved market conditions (lower than anticipated balancing market bids, for example, National Grid keeps a smaller proportion of savings.

much is likely to be made available in the BM, and whether cost minimisation might justify the use of its own STOR reserves.

79. National Grid had not forecast quantitatively the impact of behavioural change induced by EBSCR changes on its decisions in this process. However, it considered that the EBSCR would be likely to have the following effects:

(a) There might be a change in its anticipated annual reserve requirement due to changes in the physical availability of plant on the system, as described in 78(a)(i)(i) above; National Grid considered that this effect was likely to be small.

(b) Given an overall reserve requirement, less of it might be purchased through STOR because of behavioural changes induced by EBSCR (step 78(c) above).

80. The account of the impact of EBSCR on 'balancing efficiency' described by National Grid accords with the view presented in our simplified conceptual model. National Grid will see behavioural changes due to EBSCR largely in the 'lengthening' of the balancing market. More flexible capacity will have been built; this will make conditions in the near-term market more benign as well as making the balancing bids more attractive in many periods.

81. However, it is worth emphasising that National Grid anticipates that the impacts will be small. Current STOR purchases above the minimum safe requirement are only 500MW. Therefore, any reduction of this quantity through the mechanisms described is likely to be a relatively minor matter.

82. In summary, therefore, we agree with Ofgem that behavioural changes induced by EBSCR will reduce the extent of intervention by the SO and in that sense contribute to 'balancing efficiency' in the short term. However, overall system efficiency may be reduced. Perhaps the simplest way to develop an intuition behind the idea that minimising balancing costs and minimising system costs might conflict is the following: if imbalance were punished in some extreme way, firms would invest to the hilt to avoid it, thus adding to system costs. Our argument here is that RSP involves some degree of 'punitive' imbalance charging – in the sense that it is not justified on the grounds of efficient economic prices – and that this will reduce balancing costs to some degree. We have not seen an argument suggesting that this trade-off has been set in the right way. We have presented arguments

suggesting that the increment in balancing efficiency, the benefit of this policy, might be small.[46]

Price signals for DSR and imports

83. In some ways, DSR and imports might be thought of as simply additional ways to supply flexibility, and so the arguments above would apply to them also. This would the case if DSR and imports could be procured in the market like any other source of supply (or demand) and could bid into balancing and STOR-like tenders.

84. Despite this, Ofgem seems to put particular weight on the question of what the price of spot electricity should be from the point of view of encouraging the right use of these sources (see paragraph 73 above). It is clearly the case that spot prices will determine the pattern of use of flexible resources like DSR and imports. The question is whether RSP provides the right pattern of usage. If we consider a period in which a demand reduction would be attractive at the RSP-determined price but not at the short-run incremental cost of balancing, then, from the perspective of efficient resource-allocation, the DSR ought not to be used. In that sense, RSP may encourage too much DSR.

85. The right incentive for the use of a given resource at any given time is the incremental cost of the next best alternative. However, that does not necessarily encourage the right level of investment in a resource because of the problem of recovering fixed costs. This is a common problem in markets, and in some ways it is easier to solve well in highly regulated markets than in less regulated markets. The standard second best result for fixed-cost recovery is that this should be done at times when it has least impact on allocation decisions – at times of lowest elasticity of demand. Interestingly, DSR is used precisely at times of higher than usual demand elasticity, and so there is an argument to suggest that the time to recover the fixed costs of DSR are precisely not the times at which it is used. This conclusion would tend to go against Ofgem's desire to increase prices through RSP in order to provide the right incentives for DSR.

86. Incentives for the use of interconnectors are a difficult matter, especially because it is not clear what the framework of reference of welfare measurement ought to be. Should we be thinking of GB system cost

[46] The Ofgem modelling of quantitative benefits does not try to explicitly model the 'balancing efficiency' gain that we have identified as a genuine potential benefit of the reform. See Baringa Report, p10 'Although the CM will be the main tool for ensuring capacity adequacy, cash-out reform should also increase security of supply through increasing the value of flexibility. This should increase incentives to invest in flexible generation and demand side response. However, neither of these potential benefits is quantified explicitly in the modelling.'

minimisation or of EU-wide costs? If the latter, then it is hard to see that the incentivisation of imports should depart from treating imports as just another potential source of flexible generation. Ofgem has faced this question when considering (and rejecting) a modification proposal, P201, which sought the elimination of a source of incentives to importers in the method of charging BSUoS.[47]

87. In summary, it is not clear that there are special features of DSR and imports that should put them into a separate category of benefits or incentives when considering the impacts of RSP on balancing efficiency.

88. Utilita and Ecotricty make a related criticism of the EBSCR, arguing that 'adherence to marginal cost pricing cannot be justified […] where the suppliers impacted are unable to respond.' The argument is that where suppliers are settled on average consumer profiles rather than actual consumption, DSR as a response to high cash-out prices makes no sense. We sympathise with this argument, although it is too extreme as stated. Parties do have options apart from DSR in insuring themselves against high cash-out prices, like contracting for more flexible capacity, and some customers metered and settled on a half-hourly basis (currently the larger industrial and commercial customers and some smaller business customers) will be able to respond. The price signals are thus not useless. However, we agree with Utilita and Ofgem on the importance of improving system settlement in order to deliver the flexibility benefits of DSR.

Interaction between the Capacity Market and the cash-out rules

89. The EBSCR was originally intended as a set of reforms to improve incentives for investment and to counter the 'missing money' market failure. Ofgem carried out a first impact assessment that predicted that the reforms would improve capacity investment incentives.[48] After that work was completed, DECC implemented the CM[49] which effectively solved most of the missing money problem in a different way from EBSCR's. Ofgem re-did an impact assessment[50] of the reforms and concluded that they remained beneficial despite their original goal being largely superseded.

90. In this section, we examine the links between the two sets of reforms.

---

[47] See Ofgem Decision and its reasoning of its duties under EU law at p9.
[48] See Ofgem (2013), *Electricity Balancing Significant Code Review – Draft Policy Decision Impact Assessment*.
[49] The CM is described in greater detail in Appendix 5.3: Capacity.
[50] Ofgem (2014), *Electricity Balancing Significant Code Review: Impact Assessment for Final Policy Decision*.

The 'missing money problem' in 'energy-only' electricity markets

Definition

91. Revenues to generators may come either:

(a) almost exclusively from sales of energy ('energy-only markets'); or

(b) from two distinct sources of revenues: that is, revenues from sales of energy and revenues paid to generators for making capacity available regardless of actual delivery (ie electricity and capacity markets).

In Great Britain, the NETA/BETTA market was originally designed as an 'energy-only' electricity market. Internationally, we see both energy-only markets (eg ERCOT in Texas and NordPool in Scandinavia) and markets with capacity mechanisms (eg PJM in the north-eastern United States).

92. The theory of a well-functioning competitive energy-only electricity market[51] is that generators will fully recover sunk capital costs (eg the costs to build generation capacity) at very occasional peak times[52] – once every 20 years,[53] perhaps.[54] At these times, demand is high enough to give the owners of this generation capacity the ability to produce electricity (or actively to interrupt its demand) to earn a price far in excess of short-run marginal cost.

93. The theory of energy-only markets is that the promise of very occasional, very high rents in periods of extreme demand is sufficient to reward and incentivise the owners of generation capacity. Further, the additional capacity, which meets demand at very occasional peak times, provides sufficient capacity margin for the SO to balance the system safely and efficiently in the normal course of events.

---

[51] RE Bohn, MC Caramanis, FC Schweppe (1984), Optimal pricing in electrical networks over space and time, *RAND Journal of Economics* 15(3), pp360–76.
[52] For peaking plant, the only opportunities to recover sunk costs are in such periods. For other plant, some contribution to sunk costs will come in 'ordinary' periods when there are plant with higher operating costs setting market price.
[53] Twenty years is used as an example. In traditional, centrally planned electricity systems, engineers would often define adequacy standards in terms of being able to withstand a 'once in 20 years' winter'. It should be emphasised that in real (rather than theoretical) electricity markets, peaking plant can earn revenues at other times, for example by supplying essential system stability services unrelated to energy supply.
[54] If there is demand responsiveness, then voluntary demand reductions will also occasionally – and possibly much less rarely – lead to prices being set above short-run marginal cost.

94. In practice, there is considerable doubt that events in the electricity market would ever unfold quite as the theory requires.

95. Extreme demand periods in Great Britain are most likely to be in a cold winter when weather amounts to a national emergency and when high demand is compounded by supply outages.[55] There is a critique of the theory of energy-only markets that energy companies would plausibly not believe that they would be allowed to charge extreme prices in these extreme circumstances; they might not even wish to, given the damage to reputation that the appearance of such 'profiteering' would cause.

96. But if owners of generation capacity, especially peak capacity, do not charge extreme prices in extreme demand periods, and if they are competing fiercely on price at other times, then they are unlikely to recover sunk capital costs fully. Therefore, continues the critique of energy-only markets, they cannot be expected to invest adequately to provide sufficient capacity margin for the SO to balance the system safely and efficiently. This is the phenomenon widely known in this industry as the 'missing money problem'.

97. Notwithstanding the possible missing money problem, Great Britain witnessed a considerable amount of new investment in CCGT in the early years of the 21st century. However, the NETA/BETTA system has been in existence for a short period of time relative to the expected frequency of extreme events, and the system has never been tested in terms of extreme conditions (and therefore potential for extreme prices). We therefore do not know whether investors could, within the system, have hoped to recover sunk costs. Moreover, it is not clear that the system in its early days was sufficiently competitive to engender a missing money problem: less than competitive prices in ordinary times could allow market participants to recover fixed costs even in the presence of a missing money problem.

98. We have seen some direct evidence from company corporate documents that at least some generators believe that there has been a missing money problem. For example, a paper presented to the SSE board in 2013 included a comment that prices may not rise despite tightening system margins because of fear of rent extraction at peak times:

> A key risk in this assumption is the likelihood of system shortages duly materialising – but failing to translate into higher distress

[55] Alternatively, prices could rise to extreme levels in periods of ordinary demand when a large-scale supply outage had occurred: for example, a nuclear shutdown. A sufficiently catastrophic supply-side event would probably also count as a national emergency.

> prices. The experience of December [2012] is that even the last CCGT on was reluctant to extract sufficient rent to make a meaningful contribution to its fixed costs of remaining open. Owners of CCGTs are nearly all vertically integrated utilities with a cautious approach to regulatory obligations and interpretation.

99. It is plausible that investments in capacity might be inadequate if potential investors share the view that the missing money problem is material.

100. As policy to decarbonise electricity production developed in the late 2000's, it became clear that investors in thermal generation would have increasing challenges in recovering sunk capital costs. Low carbon generation mostly has very low short-run marginal costs. In an energy-only market, increased renewable capacity brought on to the system through subsidy is likely to make thermal generators more and more reliant on increasingly infrequent periods of system stress to earn a positive margin. The falls in peak demand due both to recession and to energy efficiency measures have exacerbated the problem for investors in thermal plant and for those with sunk costs that have not yet been recovered. A missing money problem has therefore become a more and more significant concern.

DECC's Capacity Market and Ofgem's cash-out reform as reactions to the missing money problem

101. Both DECC and Ofgem initially each responded to the increased challenge of the missing money problem with reforms. DECC has addressed the problem directly with the CM, a mechanism that will, from 2018, directly remunerate capacity for being available, regardless of energy produced.

102. Ofgem developed the reform to cash-out arrangements through the EBSCR,[56] and initially justified the reform in terms of increasing investment incentives. The RSP element of the reform, described in paragraph 63, above, is particularly relevant to this issue (see paragraphs 104 to 106, below). Although adequate capacity investment is no longer the primary goal of the reform, one of the four aims of the EBSCR remains to 'incentivise an efficient level of security of supply'.[57]

103. The CM is discussed in more detail in Appendix: 5.3. We believe that it is likely, and we assume for the purposes of this paper that DECC's CM

[56] Ofgem lists another three high-level objectives of the EBSCR: to improve the efficiency of balancing the system; to ensure compliance with EU rules; and to complement the DECC CM.
[57] Ofgem (2014), *Electricity balancing significant code review – final policy decision*, paragraph 1.7.

addresses any missing money problem of the sort described above as it is designed to provide payments for adequate capacity availability.

RSP and adequate investment

104. The move to **RSP** has a clear investment incentive property. STOR capacity can be thought of as being, among other things,[58] the last available increment of capacity on the short-run supply curve (see Figure 1 above). Any capacity purchased by the SO under STOR contracts can be used to supply the energy market in a 'STOR window' (a period when STOR is used). Therefore, if STOR is used *for energy balancing*, it is almost certain that suppliers and generators as a whole will have been short: some energy is needed that could not have been contracted for in the open market.[59] This is why this element of the reform will have a clear impact on wholesale electricity prices. There are periods when STOR contracts will be expected to be used for energy balancing, so the forward electricity price for those periods (or for longer periods expected to include some of these periods with some frequency) will reflect any impact of RSP on prices.[60]

105. If supply and demand are in a configuration in which parties can be almost sure that STOR will be used for energy balancing purposes, then the likelihood is high that cash-out prices will be set at LOLP × VoLL (ie above short-run marginal cost) for some of their energy.[61] When price is set outside RSP periods, there is spare capacity on the system and there is no reason for market participants to be systematically short. Competition to supply energy therefore leads to prices set at short-run marginal cost whether that is in the BM or in the bilateral market ahead of gate closure. But when STOR is likely to be used for energy balancing someone will almost certainly be left short. RSP changes the incentives around the remuneration of incremental capacity in such cases.

106. A useful way of thinking about what RSP is doing is that it is committing the owner of the last units of peak capacity – the SO under its STOR contracts – to add a margin over avoidable costs (as a monopolist would) rather than pricing at short-run incremental cost (as would a competitive market facing price competition in a homogenous good without capacity constraints). If

---

[58] As SSE and Ofgem pointed out in their responses to our market rules working paper and, STOR is used outside periods of system stress; in such periods, it is unlikely that RSP will set cash-out prices because LOLP will be very low. However, it remains true that in periods of substantial system stress, STOR is likely to be used.
[59] Strictly speaking, there is nothing to stop the SO from using STOR for energy balancing actions in a period when the system is overall long if that is the efficient thing to do. But these will be rare cases and do not invalidate the general point about the system when STOR is used as the last available increment of supply.
[60] One useful way to conceptualise STOR and RSP is to consider STOR as the last increment of capacity in the merit order; RSP is then the rule which determines how that capacity is offered to the market.
[61] Any one party may still be long, but in aggregate we know that they cannot be when STOR is used.

STOR were priced at utilisation cost and demand curtailment were not priced at VoLL, the missing money problem would be institutionalised. When demand is such that STOR must be used for energy balancing, the current system (before the introduction of RSP) ensures that the price of the last units of electricity is determined by utilisation cost, which is exactly what the theory of energy-only markets says should not happen for peak capacity increments owned by the marginal owner.

Is RSP effective in solving the missing money problem?

107. RSP satisfies some of the characteristics needed to solve the missing money problem:

(a) National Grid is committed by RSP to supplying energy at times of low capacity margin to those who are short at a price above utilisation cost, which is a requirement for the proper functioning of an energy-only market.

(b) The pricing formula based on LOLP will spread out earnings over time[62] and will not rely on a small number of extremely infrequent high prices; this may reduce uncertainty of cash flows (ie risk) for investors in capacity and should therefore lower their capital costs relative to a 'pure' energy-only market.

108. However, it is not clear that RSP constitutes a good solution to the missing money problem. There are four levels of criticism of RSP:

(a) alternative solutions to the problem may be preferable;

(b) aspects of the ESBCR reform may render it ineffective;

(c) it may be poorly designed in its detail; and

(d) it creates the possibility, together with the CM, of overpayment for capacity.

We assess these four of criticism of RSP in turn below.

Alternative solutions to the problem may be preferable

109. The CM is an alternative solution to the missing money problem. In its essence the difference between the two solutions is whether the government

[62] The LOLP formula effectively makes prices above system short-run marginal cost more frequent but less extreme than would a pure energy-only market, hence smoothing revenues of peak capacity owners over time.

chooses the level of capacity required (the CM) or the price that capacity will earn (RSP). Different international systems have chosen different approaches and there appears to be little consensus as to which is the better solution to the missing money problem.

RSP may be poorly designed in its detail

110. Even if RSP did work perfectly, the value of lost load that has been used in the RSP and forced demand reduction modifications is of £6,000/MWh (moving up from £3,000/MWh as part of transitional arrangements). This is well below the level of £17,000/MWh that DECC uses to determine the level of capacity demand in the CM. If DECC is right in this value, then the EBSCR in some ways can be seen as institutionalising missing money in that it sets a level for the most extreme prices that is too low to adequately reward the right level of investment. RWE noted that the situation with respect to security standards was even more confused than this, with National Grid making purchases of strategic reserves which resulted in a security standard of 0.6 hours of annual lost load[63] as opposed to the 3 hours implied by the DECC standard.

111. Ofgem responded to this criticism by offering three reasons for the £6000/MWh choice:

(a) It was sufficient to incentivise most industrial and commercial customers to sign-up to DSR contracts.

(b) It is a sufficient signal to improve the efficiency of interconnector flows.

(c) It provided sufficient incentives for self-balancing but was also 'designed to protect market participants from an 'unlucky day' where a VoLL of £17,000/MWh could cause significant financial distress.'

112. All but the last of these would apply also to a VoLL that is consistent with DECC's or even National Grid's higher implicit value, so we consider the last of these to be the most material. As we have seen above, the importance of self-balancing incentives needs to be judged in the context of the natural monopoly regulation of National Grid. The setting of VoLL for purposes of RSP does not relate to short-run operational incentives and short-run cost recovery, and Ofgem is arguing that VoLL is being set to insure some amount of self-balancing while acknowledging that more self-balancing, which could be induced by a DECC-consistent VoLL, would have bad consequences for

[63] See National Grid's winter outlook, p60, which says that its SBR purchases for the last winter reduced the Loss of Load Expectation from 1.6 hours to 0.6 hours.

players suffering an 'unlucky day' event. Ofgem thus appears to be arguing that one cost of this form of natural monopoly regulation is that it imposes financial risk on other parties – maybe disproportionately on smaller parties who have less sophisticated forecasting abilities and less market access – and that this cost should not be too severe. In particular, it should not be disproportionate to the benefits of 'balancing efficiency'.

113. We agree with this line of argument, and we have not seen a quantitative or qualitative arguments for why this should imply a VoLL of £6000/MWh or more or less. Our view that the impact of RSP on National Grid's STOR-purchasing behaviour is likely to be small might suggest that the value of incremental self-balancing is likely to be small too, and therefore that RSP should not aim to increase the incentives for self-balancing very much.

RSP, together with the CM, creates the possibility of overpayment for capacity

114. Ofgem and DECC have both stated that the CM and the EBSCR in general are complementary. The argument is that the contribution of EBSCR reforms towards solving a missing money problem is that bidders in the CM will anticipate these potential additional revenues, displacing revenues they would otherwise seek through the CM. This should lower the clearing price for all capacity in the CM, and lower prices would then be passed through to consumers' bills. In the extreme case in which the EBSCR reforms are believed to solve any missing money problem completely, prices in the CM should fall to zero.

115. Within the context of its assessment of the CM reform under state aid rules, the European Commission received a submission raising concerns regarding overcompensation caused by the coexistence of the CM and payments under STOR. In response to this submission, the UK government noted that capacity providers could not benefit from both long-term STOR contracts and CM contracts, and that concerns regarding overcompensation would not be present in the annual STOR auctions. This is because the STOR auction for annual contracts occurs after the CM auction has taken place, and therefore providers would be able to factor their CM revenues before bidding in the annual STOR auctions, resulting in no overcompensation. The European Commission accepted that the CM had been designed to be consistent with the reform of electricity cash-out arrangements.[64]

[64] European Commission (2014), *Letter to the UK government in relation to State Aid S.A.35980 (2014/N-2) – United Kingdom – electricity market reform – Capacity Market*, paragraph 131.

116. Under this optimistic view, RSP and the CM are offsetting, so that any additional revenue that generating capacity earns through RSP leads to a reduction in the revenue it receives through the CM. Ofgem, in its final decision, points to the possibility that the two mechanisms together might lead to a transfer to consumers from the owners of inflexible capacity.[65] However, it is not clear that this effect leads to the right incentives for investment in inflexible plant: to the extent that revenues are lowered for inflexible plant, the interaction may lead to the wrong signals for investment in flexibility.[66] The size of this transfer to consumers is likely to be modest, since the market for exclusively flexible energy and capacity is small. Hence, revenues that accrue *only* to flexible plant due to RSP are likely to be modest.[67]

117. Ofgem argued that increased revenues due to EBSCR being offset against CM bids is a good thing not only from the point of view of system efficiency but also because it means that the CM could be a temporary mechanism, to be replaced when the electricity market has settled down in its new, decarbonised configuration, with a return to an energy-only market augmented by the EBSCR. However, the CM is expected to continue to operate well into the next decade, so any benefits from easing a possible eventual transition away from the CM ought to be balanced against possible costs of having both systems working in parallel.

118. The greatest of these costs is the risk that the two mechanisms might come to generate overpayments for capacity rather than offsetting payments. If generators believe that RSP prices might not come to be allowed due to regulatory intervention, they are likely to discount anticipated revenues from RSP when determining their bids into the CM.[68] But if the mechanism then turns out to allow high prices and payments, then there will have been an overpayment: the higher price for capacity in the CM auction together with the high prices allowed by RSP. The CM has the ability to offer 15-year contracts. Thus, even if parties eventually learn to trust RSP payments and do eventually discount them from CM bids, there could still be substantial

---

[65] The argument suggested is the following: flexible plant can participate in some last-minute markets in which inflexible plant cannot participate; RSP increases revenues in those markets (as well as in other markets); so flexible plant will discount larger energy market revenues than inflexible plant in preparing their CM bids; since flexible plant are likely to be price-setters in the CM, they will set a lower clearing price and this will be a consumer benefit compared to what would have been the case without the contribution to solving the missing money problem that RSP represents. This argument points to a real net consumer benefit.

[66] Inflexible plant is already ruled out of near-term markets, and this may be a sufficient signal for investment in flexibility. The transfer to consumers should be considered a net benefit only if the incentive to provide flexibility was too low in the absence of RSP.

[67] This is a similar point to the one made in paragraphs 73 to 81 about the likelihood of a small flexibility benefit from RSP.

[68] It might be thought that the possibility of discounting higher RSP payments will be all the greater given the slightly unorthodox mechanism by which generators can turn the higher cash-out prices created by RSP into revenues.

ongoing costs from earlier rounds of CM auctions in which prices were set too high because of lack of belief in the proper operation of the system.[69]

119. [✂]. All of these examples suggest modelling methodologies that assume that missing money is addressed somehow in energy markets.

120. On the assumption that these methodologies are widespread, we would expect that, to some extent at least, CM bids will have been reduced as hypothesised in the Ofgem impact assessment. While the risk of double payment is not excluded, we have not found strong evidence that it will occur.

121. However, if there is overpayment – or even just the appearance of it – the commitment to maintain the system as it is will be made harder. This reduces the degree to which RSP contributes to solving the missing money problem.[70] The scenario that might cause concern is the following: after a period of tight operation in the market in which RSP sets prices, a question could come to prominence about why such high prices are being charged and why high operating profits are being made. We have seen that a succinct answer is rather difficult: although we see a low risk of harm, we also see no substantial benefits from RSP. Justifying the policy at times of apparent harm (very high wholesale prices at peak times) may therefore be a challenge. The degree to which the policy is believed to be sustainable might therefore come to be questioned, which increases the risk of overpayment.

[69] The opposite worry is also present: that the interaction of the two mechanisms leads to under-remuneration of capacity. Imagine that generators are overly optimistic about the degree to which RSP solves the missing money problem. They will then discount their CM bids too much and eventually lose money. This would eventually raise the cost of capital of investment and lead to higher consumer prices. The fundamental issue is that by having two overlapping mechanisms to solve missing money, an avoidable source of uncertainty is introduced into investment decisions.

[70] Some of the US markets have both CMs and RSP-like provisions. However, there are mechanisms in these markets to adjust CM payments downwards in view of high RSP payments. This sort of provision avoids overpayment risks in these designs.

Annex A: Incentives for investment in flexibility under different cash-out rules

The model

1. A supplier forecasts its net imbalance 240 minutes before a given production period. We consider its behaviour if it forecasts that it will be short. It can either buy some flexible capacity – we can think of this as an incentive for the market to supply flexible capacity – or it can wait and see what happens. During the actual production period, the system discovers its actual net imbalance. If the system is short, then the SO buys and uses STOR. The forecast net imbalance at 240 minutes is a random variable and the actual imbalance at the point of production is another independent random variable.

2. The key idea the model is constructed to explore is that contracting early (acting on the first random variable) can protect a party from imbalance charges, while contracting later, after gate-closure with the SO, uses more up-to-date information but exposes the party to the imbalance regime.

3. Our goal is to assess system performance under two different imbalance regimes:

   (a) The first is one in which suppliers have been highly incentivised to avoid cash-out and therefore purchase flexible capacity based on the anticipation of imbalance – we call this regime INVEST.

   (b) The second regime is one in which the supplier does not act on the early information anticipating a shortfall and instead pays a cash-out price based on the utilisation price of STOR, which we assume to be competitively tendered – we call this regime SO.

4. We assess system performance under two criteria:

   (a) Overall cost minimisation.

   (b) 'Balancing Efficiency', which we take to mean the minimisation of the role of the SO.

Model assumptions

5. We make the following assumptions:

   (a) Flexible capacity costs the same amount to acquire and to operate whether it is acquired in the bilateral market or in the balancing and STOR markets.

(b) If bilateral flexible capacity has been acquired and turns out to be required to balance the system, then it is used in preference to balancing or STOR options.

6. The first assumption is important. It effectively says that where flexibility can be built into the system – for example in engineering decisions on plant that are being built to provide adequate capacity – it does not matter whether the incentive comes from purchases in the bilateral market or from sales in the BM or STOR. In all cases, the nature of the flexibility investment and technology is the same. This assumption is a simplification. It is possible that some options are made available by the existence of STOR contracts and would not be available if sold only on the vagaries of the bilateral market. A diesel farm might find investment backing if its business model is STOR-based but not if it is bilateral contract-based. There is the possibility that such details do matter to the mix. However, we would not expect them to matter a huge amount – very large differences in the projects made available under the different regimes would encourage compensating changes.

7. The second assumption is natural and innocuous – it simply follows the logic of bilateral contracting and balancing.

Results

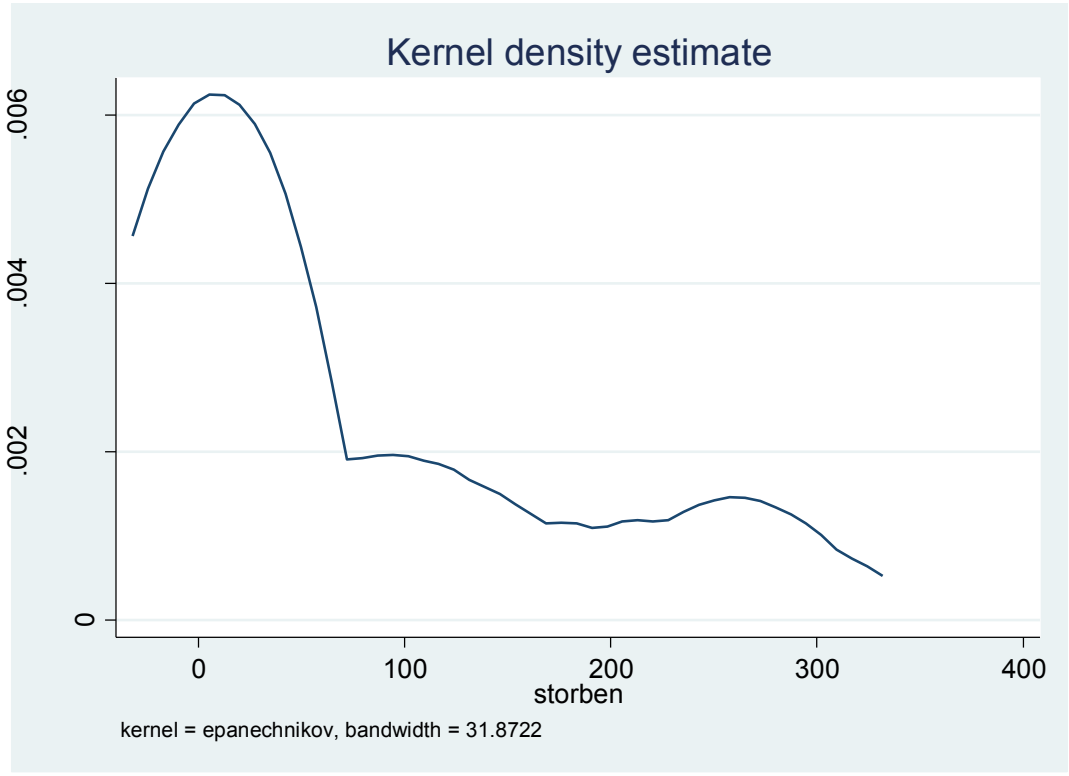
8. Under fairly generic assumptions (specifically, that the capital cost of flexibility is the same whether it is purchased in bilateral markets before gate closure or by the SO after gate closure), we find that:

(a) SO is always overall cheaper than (or equal to) INVEST (Figure 1);

(b) INVEST always has higher (or equal) investment levels than SO (Figure 2); and

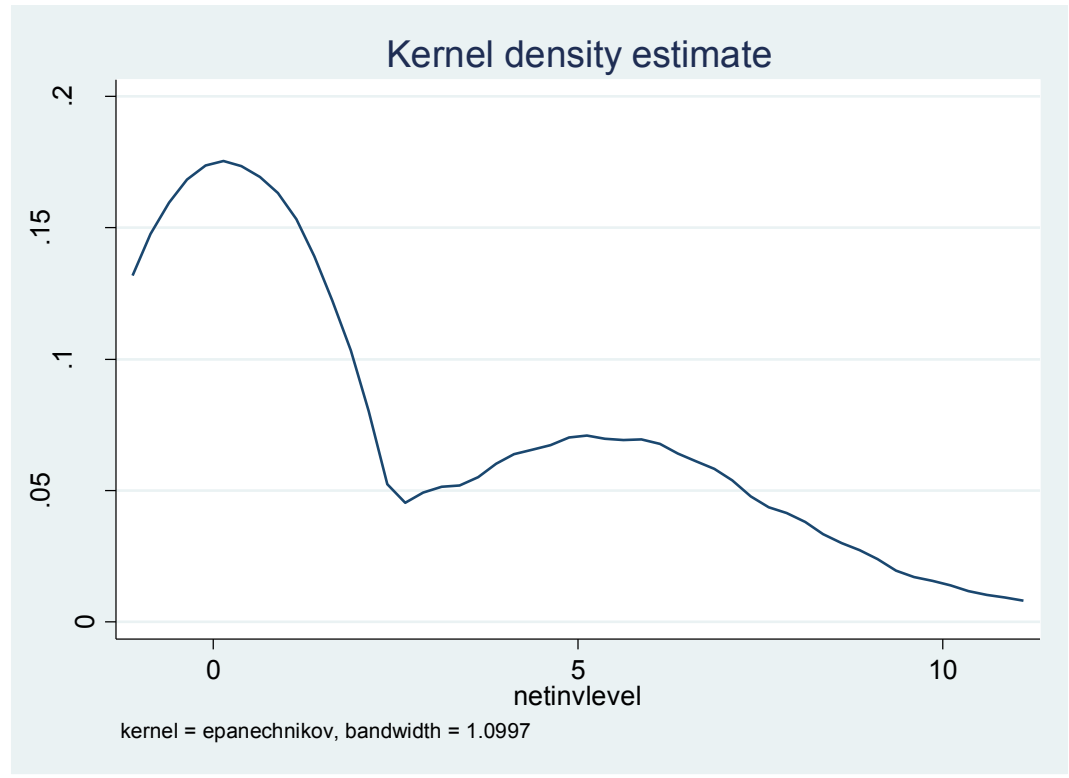
(c) SO always has higher (or equal) levels of usage of STOR (Figure 3).

Figure 1: Distribution of net benefits of SO vs INVEST system costs. Monte Carlo simulation



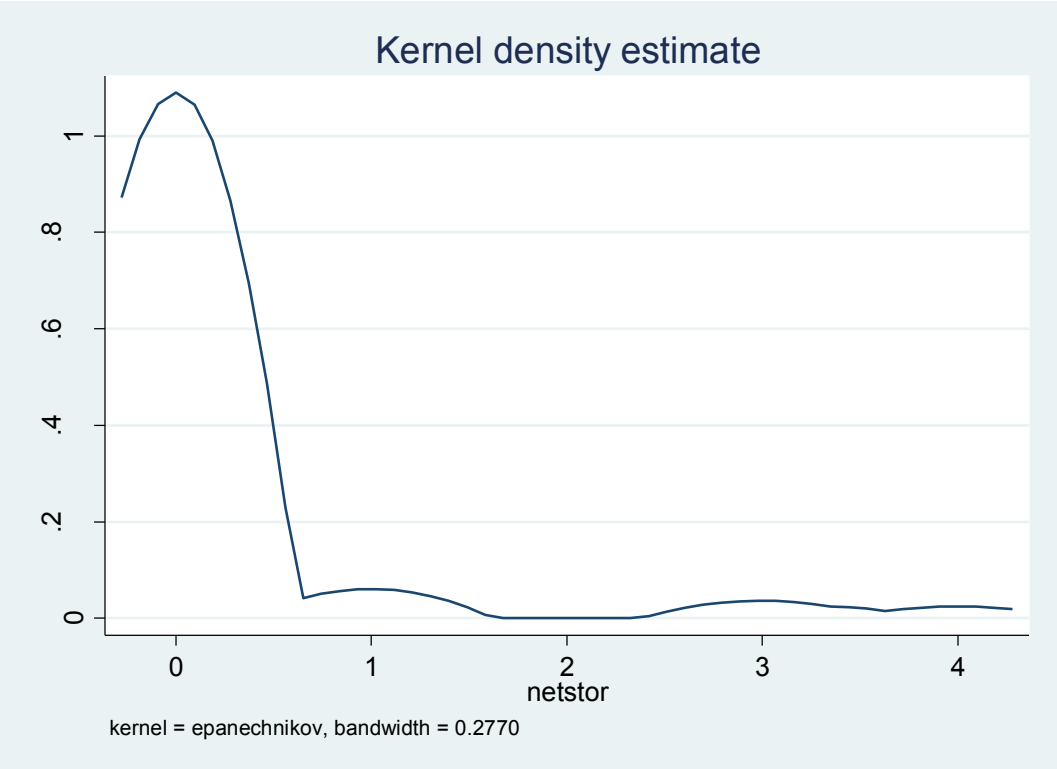
Source: CMA simulation model.
Note: The numerical values for the net benefit on the x axis are purely illustrative.

Figure 2: Distribution of net investment levels, INVEST vs SO. Monte Carlo simulation



Source: CMA simulation model.
Note: The numerical values for the investment level on the x axis are purely illustrative.

Figure 3: STOR level usage in SO minus STOR usage in INVEST



Source: CMA simulation model.
Note: The capacity levels on the x axis are purely illustrative.

9. The intuition behind these results is straightforward: by using the earlier information to make investment decisions, the supplier sometimes sends a signal to the market for flexible capacity which turns out not to be needed. There are cases in which, had the supplier waited, the imbalance might have resolved itself.[71] The SO under the SO regime, on the other hand, does not over-invest in such cases.

10. There are other cases where the anticipated imbalance is smaller than the actual imbalance. The BM is then used in both the SO and INVEST regimes, with more use in SO. However, the total amount of flexible capacity used is the same under SO and INVEST. The investment incentive is of the same magnitude, but directed differently: to market-based flexible capacity under INVEST and to SO-purchased solutions – balancing and STOR – in the two regimes.

11. System supply costs under the INVEST regime can never be cheaper than the SO and can sometimes be more expensive. However, in those cases in which the SO has to rely on STOR and in which the supplier anticipated a

[71] There are also cases in which the imbalance gets worse, so that the bilaterally contracted capacity helps to settle imbalance but needs to be complemented by SO actions.

shortfall, the system is balanced with less use of STOR and balancing under INVEST than under SO. Therefore, INVEST is more expensive (less efficient) than the SO regime but the SO regime has a larger role for the System Operator.

12. In this sense, the model seems to capture the essence of the trade-off between the two objectives. If we believe that STOR procurement is inefficient, we will be prepared to sacrifice theoretical cost-minimisation for balancing efficiency. However, if STOR procurement is relatively efficient, then INVEST leads to over-investment in flexible capacity, in forecasting and in reliability.

13. We take this to be a reasonable conceptual model of RSP in relation to flexibility incentives. If RSP is solving a flexibility problem, it is a problem relating to the efficiency with which the SO procures flexibility services. The market failure is a regulatory failure.

14. The model suggest alternatives methods to reducing the role of the SO that have a lower risk of leading to over-investment in flexible capacity or of leading to double payments. If the supplier can trade closer to the delivery window, the need for non-traded STOR purchasing is reduced. Moving gate closure closer to real time would seem like a simple way of encouraging the substitution of open market investments in lieu of SO-purchased STOR.