2. Wholesale electricity markets

Guido Cervigni and Dmitri Perekhodtsev

2.1 INTRODUCTION

Trading electricity in a similar way to other commodities requires special arrangements. Transaction costs are reduced by product standardisation, and a central agent, the system operator, ensures that production and consumption match on a constant basis regardless of actions by generators and consumers. Figure 2.1 illustrates the typical timeline of electricity transactions.

In this chapter we discuss transactions taking place in wholesale electricity markets, splitting them into two groups: electricity market transactions and transactions related to system operations.

2.1.1 Electricity Market Transactions

This group includes transactions taking place between market participants up until shortly before the actual time production and consumption take place, which is referred to as the ‘real time’. We refer to the broad range of such transactions as ‘market transactions’.

All the market transactions are forward transactions, because they take place before the time of delivery. They are concluded between years

Figure 2.1 Timeline of electricity transactions

or months and a few hours ahead of real time. Products exchanged in market transactions are often highly standardised, both in terms of time and space. For example, a typical traded product is a volume of electricity produced and consumed in a given hour at any location in a given country. Production or consumption taking place at different time intervals within the hour or at a different location within the same national network are therefore considered identical products.

The arrangements governing forward transactions taking place near the time of delivery, generally starting from those taking place the day before delivery, are specifically designed to address the high volatility of electricity demand and supply. The price of electricity can differ dramatically from one hour to the next depending on the level of demand and on available generation capacity. In addition, the expected demand and supply conditions at a certain time of delivery may also change dramatically as the time of delivery approaches. In this context, decentralised or bilateral trading would entail high transaction costs and could lead to large-scale inefficiency. Organised markets, or power exchanges, reduce transaction costs and ensure an efficient market outcome by centralising the negotiations.

The latest time when market transactions for delivery at a certain time can be entered is called 'gate closure'. After gate closure the market participants inform the system operator which generating units they intend to activate in order to produce the electricity sold on the market and where in the network the electricity bought on the market will be consumed.

2.1.2 System Operations

The system operator concludes transactions with market participants as part of the activities it carries out to ensure that all the system's security constraints are constantly met.

Transactions related to system operations take place at different times around the moment of delivery. Starting from gate closure and through the time of delivery, the system operator procures additional production or reduced production on the real-time or balancing market. In the balancing market the system operator sells and buys energy from generating units and consumers with load-control capability, to match instantaneous imbalances between production and consumption. In order to ensure that there is sufficiently flexible capacity to perform balancing in real time, the system operator may procure reserve capacity in advance of the gate closure. Finally, both before and after gate closure, the system operator carries out actions intended to ensure that production and consumption

can be securely supported by the transmission network capacity. We refer to these actions as ‘system re-dispatch’.

The system imbalance, offset by the system operator in real time, is the sum of the individual imbalances of market participants, that is, of the deviations of market participants’ actual production and consumption from the quantities they have notified to the system operator as respectively sold and purchased on the market. Once the actual consumption and production of each market participant is known, based on metering data, the system operator financially settles the imbalance of each market participant. The prices at which imbalances are settled are known as ‘imbalance charges’ and are closely related to the prices prevailing on the balancing market. The process that determines the price and the volumes of those transactions is known as an ‘imbalance settlement’.

All the transactions analysed in this chapter are highly interdependent. Most of them refer to the same product: energy produced and consumed at a given time. This is the case for the forward transactions between market participants, the real-time transactions between the system operator and the balancing energy suppliers, and the transactions that settle the market participant imbalances. In addition, some transactions involve products that are jointly produced; in particular a generator’s capacity can be used either to produce electricity or to provide reserves.

The rest of the chapter is divided as follows. In Section 2.2 we discuss market transactions. In Section 2.3 we address system operation-related transactions. In Section 2.4 we compare the philosophies underlying the electricity and ancillary service trading arrangements implemented on the European and US markets.

2.2 FORWARD TRADING

Contrary to most other goods, even small changes in demand and supply conditions have a major impact on electricity prices, since stockpiles cannot be used as a buffer. Market participants hedge against price volatility through forward trading. Electricity in the wholesale market is bought and sold between power generators producing electricity from power plants, retail suppliers serving end customers and traders that do not control generating units or serve consumers.

Electricity is traded in the forward market in different timeframes, ranging from some years to a few hours before the time of delivery. The latest moment in time when market transactions for delivery at a certain time can be entered is called ‘gate closure’; typically the gate closure is set

Figure 2.2 Timeline of electricity transactions

one or two hours before real time. Figure 2.2 illustrates the portion of the electricity transaction timeline presented in this section.

In Section 2.2.1 we analyse the arrangements governing transactions entered into the day before delivery, the day-ahead market. In Section 2.2.2 we address transactions entered into on the day of delivery, the intraday market. In Section 2.2.3 we discuss longer-term transactions. Finally, in Section 2.2.4 we illustrate how electricity transactions translate into production and consumption commitments.

2.2.1 Day-ahead Markets

We begin our discussion of electricity forward transactions from the arrangements for day-ahead trading, because they are affected more than longer-term transactions by the technical features of electricity. On the day before delivery, market participants obtain relatively reliable information about the demand and supply conditions. In addition, some thermal generating units are rather slow to start producing power and the decisions to start up, or commit, these units need to be taken in the day-ahead timeframe. The design of the day-ahead market is crucial to the system's efficiency, since its outcome determines which generators will be started up and will therefore be available to match load in real time.

Demand and supply conditions may and do change after the day-ahead market is cleared. Producers and retail suppliers can adjust to such changes by trading in the intraday markets a few hours before real time. Increased electricity generation from intermittent renewable sources the availability of which is best predicted close to real time increases the importance of intraday markets.

This section is organised as follows. First, we discuss the rationale for centralising day-ahead transactions in power exchanges or pools. Second, we discuss the merits of the much-debated non-discriminatory or marginal price auction design implemented in most day-ahead markets. Third, we

discuss pricing in conditions of scarcity. Fourth, we analyse how alternative market designs address the trade-off between the need for product standardisation on the one hand, and the need for generators to sell products that it is feasible to deliver given the technical constraints on the other.

Organisation of the day-ahead market: bilateral transactions, exchanges and pools

Much of the debate surrounding the restructuring of the electricity industry in the 1990s in both the US and Europe focused on how the electricity day-ahead market should be organised.

Three organisation schemes have received the most attention: bilateral markets, exchanges and pools. In a bilateral electricity market, buyers and sellers trade directly with no coordination by a central body. As a consequence, price discovery entails repeated one-to-one interactions between market participants.

An exchange is an entity that coordinates the trading of standard products between market participants. It collects buy bids (or simply bids) and sell bids (or offers) for electricity, and clears the market. Coordination provided by the power exchanges is crucial for efficiency in the electricity industry, because the value of electricity differs significantly from one hour to the next. Most exchanges trade hourly energy products with delivery on each hour of the day, that is, a total of 24 hourly products per day.¹ In the event that network-related constraints are enforced on market transactions, as we discuss later in Chapter 4, the value of electricity differs not only in time but also by location, and an even larger number of prices may need to be discovered by the market participants. For those reasons, discovering the value of electricity each hour via bilateral negotiations entails significant transaction costs. Exchanges reduce transaction costs by identifying the market-clearing prices and the corresponding series of transactions. However, trading outside the exchanges through bilateral negotiations is generally also allowed. Moreover, in some countries such as the UK, multiple exchanges clear the same physical market.²

When trading bilaterally or through an exchange, sellers do not make commitments as to which generating units will actually produce the electricity they have sold, and buyers do not make commitments as to where they will consume the electricity they have purchased. In contrast, pool market-clearing systems also perform scheduling. They determine the set of generation units that will be started up, as well as the production level of each unit in each hour and the hourly consumption by each market participant in each location.

Furthermore, the pool market-clearing algorithm ensures that the

market outcome is feasible and secure. Feasibility means that the market-clearing production schedule for each unit is consistent with the unit's technical capabilities. In this respect, compared with an exchange, a pool provides a higher degree of coordination of market participants' decisions. Security means that the set of market-clearing schedules (i) satisfies all the network's security constraints, and (ii) leaves an adequate reserve margin for real-time balancing.

Those features have several major implications for the design of pool-based systems, ranging from the bid format to the market-clearing algorithm, which we investigate in the following sections and in Chapter 4.

With the partial exceptions of Italy and Spain, the day-ahead trading model currently prevailing in Europe is based on bilateral trades and trades through power exchanges. Market clearing and scheduling are carried out separately. After the gate closure, each seller decides which units will produce the electricity it has sold and each buyer decides where the electricity it has purchased will be consumed. These decisions are notified to the system operator in the form of injection and withdrawal programmes. After the schedules have been notified, the system operator takes the measures necessary to ensure that the system is secure.

In the rest of the chapter we base our presentation of the wholesale electricity markets on the European model, although in Section 2.4 we compare this model with the pool-based model implemented in much of the United States.

Market participants generally enter the day-ahead stage with open trading positions taken in the forward markets. They use the day-ahead market to adjust their forward positions. However, in order to simplify the presentation of day ahead markets we have assumed that no positions have been taken by the market participants prior to the day ahead, and we have ignored speculative purchases and sales. This implies that supply in the day-ahead market corresponds with the generators' production capacity, and that demand corresponds with expected consumption. Finally, we have assumed that bids and offers in the day-ahead market do not reflect the expected outcomes of the subsequent market sessions. In other words, we have assumed that no arbitrage takes place between the day-ahead and the intraday and balancing markets. We remove these simplifying assumptions in the following sections and chapters.

The auction model: non-discriminatory versus pay-as-bid auctions

In the day-ahead trading session, market participants submit offers for production and bids for consumption of electricity on the following day. The market operator accepts bids and offers in order to maximise the net gains from trade, or the surplus generated by the transactions.

Figure 2.3 Market-clearing price in a day-ahead power exchange

For the sake of simplicity of exposition we have discussed alternative auction models in a market with only simple bids and offers. Each simple bid or offer refers to consumption or production in a single given hour.³ In a market with simple bids and offers, the set of accepted bids and offers cleared for each hour is independent from the sets in the other hours. In this case the market equilibrium for one hour can be characterised as the point where the bid-based supply and demand intersect,⁴ as shown in Figure 2.3.

The pricing and clearing rule implemented in the day-ahead markets has been the subject of extensive discussion.⁵ All day-ahead power exchanges and pools that we are aware of run non-discriminatory auctions (also known as ‘single clearing price’ or ‘single-price’ auctions). In such auctions, every accepted bid and offer, respectively, pays and receives the market-clearing price, independently of the bid and offer prices.

Standard economic theory provides the basis for applying the market-clearing price to all accepted bids and offers. Provided that the market is competitive, in a non-discriminatory auction a generator’s offer price reflects its incremental or marginal costs, while the bid price reflects the value of electricity to the consumer. The market-clearing price implements the set of transactions that maximises the gains from the trade and ensures efficient dispatch, that is, minimises total generation costs. The expected

Figure 2.4 Bidding in single-price and pay-as-bid auctions

future clearing prices will drive efficient investment decisions, that is, cost-minimising generation technologies will be selected and new capacity will be built at the right time.

The alternative auction model is known as ‘discriminatory’ or ‘pay-as-bid’. In a pay-as-bid auction, each accepted offer receives and each accepted bid pays their asking price.⁶ For the sake of simplicity, we refer here to an implementation of the discriminatory pricing rule where buyers pay a uniform price equivalent to the average price paid to sellers, so that the market operator’s budget is in balance.

Opponents to the single-price model often put forward a fairness argument. They argue that the single-price model results in excessive payments to generators, since low-cost units, such as nuclear, hydroelectric or even coal-fired power plants receive the same price as high-cost marginal units. This argument is flawed for several reasons. First, it relies on the assumption that generators submit the same offers in single-price and pay-as-bid auctions. In fact, rational generators bid differently in different auction models. This is illustrated in Figure 2.4. The left panel of the figure shows the competitive market outcome of the single-price auction. The generators’ offer prices are equivalent to the corresponding variable costs and the clearing price is the price of the highest accepted offer. In this case, revenues collected by the generators on the market are represented by the sum of the areas X and Y .

According to the naive opposition argument, if the market cleared according to the pay-as-bid model, the generators would not change their bidding strategy. In this case their total revenues would fall to the area Y in the left panel in Figure 2.4. In fact, in a pay-as-bid auction rational generators would guess the price paid to the highest accepted offer – the clearing price in the single-price auction. They would then align their offers to that

expected price. As shown in the right panel of the figure, there is no reason why generators A and B might want to sell at a price lower than P , since other generators, whose variable cost is higher than P , provide no competitive pressure to do so. If generators can correctly predict the clearing price, the total payments they receive will be approximately the same in both the single-price and the pay-as-bid auctions.

Note, incidentally, that convergence of offer prices to the market-clearing level is not restricted to pay-as-bid auctions. In ordinary markets where transactions are agreed upon on a bilateral basis, all transactions usually take place at prices that approximate the market-clearing price, despite variations in production cost among suppliers and differences in the availability to pay among consumers.

The second flaw of the argument in favour of the pay-as-bid model is that, even assuming that payments to generators are smaller compared with single-price auctions, if the market is sufficiently competitive such a reduction of the generators' revenues is not desirable. Higher gross margins for low variable-cost generators are necessary in order to remunerate the higher fixed costs of units that use less expensive fuel, such as hydroelectric, nuclear, solar and wind power. Without those margins, capital would not be attracted to electricity generation capacity and the system would eventually experience shortages.

A similar argument applies to another (alleged) virtue of pay-as-bid auctions: lower price volatility across hours. Price volatility is physiological in the wholesale electricity market, because different generators have very different variable costs and demand is volatile. It is therefore efficient for the market-clearing price to vary considerably in time. Extremely high prices when the market is tight are necessary to attract investments in peaking units, which produce a limited number of hours per year, and to induce demand to voluntarily reduce when supply is scarce. Moreover, forward and future contracts allow transfer of the price volatility risk to the parties that can bear it at the lowest cost.

In the presence of market power a reduction in the payments to generators may be desirable. Theoretical analyses⁷ find that prices are lower under pay-as-bid than under non-discriminatory auctions in simplified settings, although to a much lesser extent than the naive hope that would expect each generator to cash in only its variable cost. The limited available empirical evidence shows no significant differences in prices between pay-as-bid and non-discriminatory auctions. Some experimental studies find higher prices in pay-as-bid auctions compared with non-discriminatory auctions.

While its superiority is not proven, some features of pay-as-bid auctions make them less attractive than non-discriminatory auctions in the

electricity spot markets. First, the generators' profit-maximising strategy under pay-as-bid requires prediction of the market-clearing price. This prediction will necessarily be imperfect, and various market participants will forecast different values for the market-clearing price. As a result, a generator with a lower offer and high marginal cost will sometimes be selected rather than a generator with a lower marginal cost, and total generation costs will not be minimised.

Second, in a non-discriminatory auction the profit-maximising strategy for a competitive generator entails offering a price equivalent to the variable cost.⁸ This relatively simple competitive benchmark facilitates the detection of market power, as large deviations of offer prices from variable costs can be taken as a signal of the exercise of market power. In some electricity markets, market-power mitigation measures are built around this competitive benchmark.⁹ Specifically, market-power mitigation is obtained by administratively replacing the generator's offer prices with the estimated variable costs when offers of these generators are likely to constitute market-power abuse. In some US markets such as Pennsylvania–New Jersey–Maryland (PJM), California (CAISO) and Texas (ERCOT), automatic price-mitigation schemes are employed when offers are deemed non-competitive.

In contrast, in a pay-as-bid auction the competitive offer price systematically deviates from the variable cost. This makes the competitive benchmark more difficult to identify and the assessment of the exercise of market power highly controversial. As a consequence, under a pay-as-bid system, market monitoring would become more complex, and many of the current market-power mitigation measures would be inapplicable.

Pricing in conditions of scarcity

As long as available generation capacity is lower than demand, the clearing price of the wholesale electricity market equals the variable cost of the most expensive generator that needs to be activated to meet load, or the system marginal cost¹⁰. The wholesale market-clearing price is subject to discontinuity when demand is greater than the available generation capacity, that is, in conditions of scarcity. In this case, the competitive market-clearing price is no longer equivalent to the marginal generation cost; instead, the clearing price is the one that rations demand. We refer later to the difference between the market-clearing price and the system marginal cost as the 'scarcity rent'.

The steeper the demand curve, the bigger the difference between the price in normal conditions and the price in conditions of scarcity. Figure 2.5 illustrates this feature, based on the assumption that demand is perfectly inelastic in the relevant price range.¹¹

Figure 2.5 Scarcity rent

Currently, a large portion of electricity demand is indeed price insensitive in the day-ahead timeframe. This reflects the typical retail supply arrangements that charge the same price for the consumer's total consumption over long time periods, ranging from one to several months. Most of the electricity meters currently in place do not record consumers' hourly consumption, but only total withdrawal over a longer time period. This makes charging a different price for consumption in different hours impossible. The insensitivity of short-term demand to price also reflects consumers' preferences, as the value of electricity to consumers is generally much higher than the typical production cost.¹²

Should a price increase not reduce demand, quantity rationing must be implemented in the event of scarcity. Such scarcity may occur, for example, in the event that the day-ahead market cannot clear because the price-independent demand bid quantity is greater than the offered quantity. In this case, the system operator plans curtailment of service on different portions of the grid (rolling blackouts).¹³ Since selective disconnection is technically infeasible, all the consumers connected to the same network branch will be disconnected at the same time.

When scarcity occurs and demand is totally price inflexible, the price for electricity is set to an administratively defined value, the value of lost load (or VoLL). VoLL is intended to be the price that makes consumers indifferent between consuming electricity at that price, and not consuming. VoLL is typically estimated several orders of magnitude greater than average electricity prices.

The values of VoLL vary between countries. For example, in the UK the VoLL used is €4.18/kWh, in Italy it is €10.8 and €21.6 per kWh, respectively, for residential and business customers, in Ireland €7.2/kWh, in Norway €0.96, €11.8 and €7.9 per kWh, respectively, for residential,

commercial and industrial customers, and in Sweden €12.0, €8.8 and €7.9 per kWh, respectively, for urban, suburban and rural customers.¹⁴

Alternative spot-market designs differ in the way the clearing price is set in conditions of scarcity. One approach relies on market participants offering prices higher than their variable costs when conditions of scarcity are expected. When the system is known to be stretched, each generator calculates a certain probability that demand will be greater than the total available capacity. The generator's expected profit-maximising strategy in that situation entails offering part of its capacity at prices greater than the variable cost. By doing so, the generator takes into account the possibility that its offer will set the market-clearing price. In this situation the generator bears the risk that conditions of scarcity will actually not arise, and that its offer will be displaced by the competitors' offers. Productive inefficiencies may also arise if some offers above production cost turn out to be displaced by cheaper offers submitted by less-efficient generators. Finally, in this model the market-clearing price may turn out to be greater than the system variable cost, even if conditions of scarcity do not actually come about.

In the alternative approach, a scarcity-pricing mechanism is included in the market-clearing algorithm: the market-clearing algorithm automatically sets the prices for electricity and for the operating reserve at the VoLL when conditions of scarcity are detected based on the offers and bids submitted by market participants. In this situation, the generators' competitive bidding strategy is to offer its variable cost, irrespective of the expected demand and supply conditions, because in the event of scarcity the clearing algorithm itself will set the price at the VoLL. This approach is consistent with the broader objective of reducing the scope for inefficiencies caused by prediction errors. This requires the market participant's profit-maximising bidding strategy not to depend on its expectation of the clearing price.

In Europe this approach was followed in the pool system implemented in England and Wales between 1990 and 2001. Based on the available capacity offered in the pool and the expected demand, the system operator would assess the probability of not being able to serve the entire load the following day. Based on that probability and on the VoLL, the system operator would compute the expected value of the scarcity rent. This would then be added to the system marginal cost to determine the market-clearing price. In Nordpool, if the level of available capacity is such that the system operator must provide additional supply out of its capacity reserves, then the day-ahead price is increased to the price cap, and prices in the intraday and balancing markets must be as high or higher. Some US markets such as New York (NYISO), New England (ISO-NE), California

(CAISO) and Pennsylvania-New Jersey-Maryland (PJM) have or are currently developing scarcity pricing mechanisms.

Pricing in conditions of scarcity is a crucial element of the wholesale electricity market's design. Since the available generation capacity is far greater than demand in most hours, the competitive market-clearing price very rarely departs from the system marginal cost. Therefore the generating units with the highest variable costs rely on the extremely high prices prevailing during very few hours of scarcity to cover their fixed cost. If prices fail to reach the VoLL under conditions of scarcity, this may discourage investment in production capacity. That is a reason of particular concern in electricity, since the value of electricity for the consumers is generally much higher than the cost of generation. We address generation capacity adequacy in greater detail in Chapter 3.

Inter-temporal constraints

The cost function of most large thermal generating units reflects inter-temporal or dynamic constraints. Dynamic constraints take several forms, including (i) start-up cost: a cost borne every time the unit is brought into service; (ii) minimum technical output: a minimum level of production that the unit must deliver for technical or environmental reasons once it is brought into service; (iii) maximum ramp-rates: constraints that restrict the difference between the unit's production in one hour and the next; (iv) minimum up time: the minimum time that the unit needs to run once brought to service before it can be switched off; and (v) minimum down time: the minimum time that the unit needs to remain switched off before it can be brought into service again.

So far we have considered simple hourly bids and offers that are accepted or rejected independently for each hour. If only simple bids and offers are traded, the market outcome might result in production commitments that could not be accomplished by generating units subject to dynamic constraints.

This possibility is illustrated in Figure 2.6, where we show a hypothetical set of accepted offers in the day-ahead market for a generator that offered its entire capacity (10 MW) in each of five hours at a price equivalent to its variable cost (€50/MWh). The generating unit has a start-up time of three hours and, once activated, must produce at least 1 MW. The generator's offers were accepted in hours 1, 4 and 5 but not in hours 2 and 3. The corresponding production schedule is not feasible for the unit, since it involves zero production for two hours, while it takes three hours to bring the unit back in service after shutdown.

In this situation, the generator's production, in some hours, must depart from the quantities sold in the market. The generator selects the

Figure 2.6 Infeasible market schedule

Figure 2.7 Feasible production schedule options

lowest-cost course of action between (a) starting up the unit only in hour 4 and buying replacement power to cover its cleared sales in hour 1, and (b) starting up the unit in hour 1, running it for all the hours at least at the minimum load, and selling the excess production in hours 2 and 3. The profit in both cases will depend on the price at which the generator is able to buy or sell replacement power. Figure 2.7 illustrates the generator production in the two cases.

A market where the products traded are not consistent with the generators' cost function has two undesirable consequences. First, generators are exposed to risk. The profit-maximising offer strategy by a (competitive) generator depends not only on its cost but also on the expected market outcome. If the generator's forecast of the market outcome turns out to be incorrect, so that it has to produce at a loss or that it forgoes profitable sales,¹⁵ the generator might wish it had offered differently. Higher risk for the generators, other things being equal, increases the expected rate of return necessary to attract investment in the generation industry, and as a result leads to higher wholesale electricity prices.

Second, the fact that each generator addresses its dynamic constraints independently of the others may result in production inefficiency. When multiple generators, each with its own expectation on the cost of adjustment, independently assess their cheapest adjustment strategy, the resulting production decisions might not minimise system-wide generation costs.

Various solutions have been implemented in order to address intertemporal constraints. The first is giving generators the opportunity to make adjustment trades after day-ahead market clearing. Generators can trade bilaterally after closure of the day-ahead markets or on the intraday exchange. Some markets, such as the one in Italy, provide special adjustment sessions run immediately after the day-ahead market outcome is known, where simple hourly products can be traded again. In these market sessions the generators can modify the positions resulting from the day-ahead market through additional purchases or sales, in order to commit to a feasible production schedule.

Second, most European day-ahead exchanges allow block order trading. A block offer (or bid) is a commitment to produce (consume) a constant amount of power in a group of consecutive hours at an average price no lower (greater) than the one specified in the offer (bid). Block bids and offers are either entirely accepted or entirely rejected. A generator submitting, for example, a 12-hour block offer for its minimum technical capacity, at a price equivalent to the average generation cost including start-up cost, is sure that its unit will either operate above cost or not be activated at all. Because of the indivisibility of the block offers, the day-ahead market-clearing algorithm becomes more complex, as the impact of accepting a block bid or a block offer extends over multiple hours. Block products do not ensure that the market outcome is fully efficient or, from a different perspective, that no risk is put on the generators. Block products put the generator at risk of the chosen length of the block offer turning out not to be optimal. For example, a 12-hour block offer for a generator's minimum technical level at a price equivalent to its average cost (including

start-up cost) might be rejected, whereas an eight-hour block offer at a price equivalent to the average cost might have been accepted.¹⁶ In this case the generator offering the 12-hour block would forgo a profitable sale, and total generation costs would not be minimised.

The third approach is implemented in the pools, where generators submit offers that closely mimic their cost functions and the market-clearing algorithm explicitly takes into account the dynamic constraints of each generating unit. As a result, the market outcome is such that overall generation costs are minimised and each unit is allocated a technically feasible production programme.

2.2.2 Intraday Markets

Demand and supply conditions may change after the day-ahead market has cleared. Generating unit outages may reduce available capacity. As the time of delivery approaches, the availability of renewable sources such as wind and sunshine may change compared with the expectations on which day-ahead trading decisions were based. Electricity demand from industrial customers may vary according to the requirements of production processes, while demand from small businesses and residential customers may vary with changes in weather conditions.

Intraday markets allow market players to carry out transactions up to a few hours before real time. For example, a wind generator may become aware that a wind drop will make production of the volumes sold in the day-ahead market impossible. The generator can then honour its delivery commitment by purchasing the volumes it cannot produce on the intraday market.

Two broad models have been implemented in the intraday timeframe. The first relies on a non-discriminatory auction similar to the one clearing the day-ahead market. The clearing of the intraday market sessions occurs at regular intervals during the day of delivery. In Europe this approach is implemented in Italy and Spain. In the US standard market model non-discriminatory auctions clear the market every five or 15 minutes.

An alternative model implemented in the intraday timeframe is continuous trading. With continuous trading, market participants post offers and bids on an electronic billboard managed by the market operator. Each time a bid is submitted, its price is compared with the prices of the offers already posted and not yet matched. If one or more opportunities for positive net-value deals are available, the bid is matched with the offer that maximises the transaction's net value, that is, with the lowest-priced offer. The price for the transaction is set as the price of the offer picked from the billboard to clear the bid. If no opportunities for positive net-value deals

are available, the bid submitted is left posted in the repository. The same happens when an offer is submitted.

Support for continuous trading in the intraday timeframe comes particularly from traders, who see the price volatility as an opportunity for profitable deals around real time. Auctions might instead reduce the value for a trader of being able to access and process the information about demand and supply changes more quickly than others.

The drawbacks of continuous trading are of the same type as those discussed in Section 2.2.1 with regard to decentralised trading: the sequential matching process and the pay-as-bid pricing rule do not ensure an efficient market outcome and price discovery. Furthermore, in the current implementations of continuous trading for cross-border transactions, network-related constraints are enforced by clearing transactions on a first-come, first-served basis: bids and offers are matched across countries based on the order of arrival on the billboard until the available cross-border capacity is fully utilised. As a consequence, the set of cross-border transactions resulting from continuous trading might not be the one that makes efficient use of the transmission capacity in the event of congestion.

The intraday market implemented in Nordic countries by Nordpool gives market participants the opportunity to continuously trade hourly power products, as well as block orders. Trading takes place every day until one hour before delivery. Intraday transactions in Nordpool account for around 1 per cent of all spot transactions. Organised intraday market based on continuous trading design is also applied in the Central Western European market combining France, Germany, Belgium and the Netherlands. Furthermore, bilateral trading can be carried out in the intraday timeframe.

2.2.3 Long-term Transactions

While short-term trading is meant to ensure that electricity production costs are minimised at all times, long-term transactions play a key role in sharing the risk among market participants.

Electricity generation is highly capital intensive and highly risky. For that reason the possibility for generators to transfer risk via long-term contracts is crucial in attracting capital to the industry at minimum cost. The level of long-term contracting is commonly regarded as an indicator of the maturity of a wholesale electricity market.

Long-term contracts for electricity are traded similarly to those for most other commodities. In Europe, long-term wholesale electricity contracts are traded in over-the-counter (OTC) markets (that is, directly between

the counterparties), as well as through organised power exchanges where transactions are mainly financial and products are standardised.

The volumes of long-term contracts traded vary significantly between European countries. In 2010 the volumes of long-term contracts traded in Germany and France via power exchanges amounted to around 500 TWh.¹⁷ In Italy only physical contracts are traded, and the volumes traded over the counter and via power exchange amounted to around 310 TWh in 2010. NASDAQ OMX Commodities Europe is the most developed European forward market, allowing trading of financial derivatives contracts on the Nordic, German, Dutch and UK power markets. In 2010 there were transactions for 3,400 TWh, of which 2,100 TWh were over the counter and 1,300 TWh via power exchange.

The range of products traded includes base- and peak-load futures, forwards, options and contracts for difference. The trading time of these products ranges from a week to six years ahead of time of delivery. On power exchanges most traded products are futures contracts and options settled by financial payments rather than physical delivery.¹⁸ These transactions are highly standardised with respect to the contract specifications, trading locations, transaction requirements and settlement procedures.

2.2.4 Market Position and Physical Nomination

We have described above the sequence of transactions among market participants starting from long-term to day-ahead (D-A) and intraday (I-D) transactions. Trading in the different timeframes allows market participants to update their contract positions as the expected economic conditions at the time of delivery change.

At gate closure, trading between market participants stops. At that moment, the sum of all the contract positions of each participant with respect to the given delivery hour determines the physical obligation of that market participant in the delivery hour: a party with a net seller position has to produce a matching volume of energy from its generating resources; a party with a net buyer position is committed to consuming the matching volume of energy.¹⁹ At gate closure each market participant must submit its physical nominations, also called ‘production/consumption programmes’, to the system operator, specifying which generating units will produce the electricity to match its net seller positions, or where in the network the consumption matching its net buyer positions will take place.

Figure 2.8 is a simplified illustration of the evolution of the contract position of a hypothetical market participant with respect to a delivery hour, and conversion of the net contract position into the physical nomination at gate closure. The generator has sold 100 MWh under a

Figure 2.8 Market participant's net contract position at gate closure

multi-year contract and a further 50 MWh under an annual contract. A month before the delivery date it has the opportunity to buy 30 MWh at a price below its variable production cost. In the day-ahead market the generator buys a further 30 MWh at a price below its variable cost. On the delivery day an unexpected outage limits the generator's production capacity to 70 MWh. It then buys 20 MWh on the intraday market to make up for the reduced generation capacity. The generator's outstanding contract position at gate closure is given by the sum of all its previous positions and is equal to $100 + 50 - 30 - 30 - 20 = 70$ MWh. At this point the generator is required to match the 70 MWh contract position by notifying 70 MWh of production.²⁰ The balance of all the contract positions within a system is zero at any time, because each buying position has a matching selling position. Furthermore, since each market participant must notify a physical position that matches its final contract position, the balance of all the physical volumes notified at gate closure within a system is also zero, for each delivery hour. The following simple example provides an illustration of this.

Consider three market players. The first one is a utility that owns two generating units, Gen A and Gen B, and supplies two consumers, Cons 1 and Cons 2. The second one is a retail supplier that serves one consumer, Cons 3, but has no generation capacity. The third one is a trader that does not have any generation capacity or supply any consumers. Assume that

Figure 2.9 System net contract position at gate closure

the following set of transactions take place between the market participants. The utility sells 50 MWh to the trader; the latter sells 50 MWh to the retail supplier. As a result of these two contracts, at gate closure the utility has a net contract position of -50 MWh, the retail supplier of +50 MWh and the trader's position is zero. The combined contract position of the system is zero, since the positions of the three players are matched.²¹ This is illustrated in Figure 2.9.

After gate closure, each market participant notifies the system operator of its physical positions. The balance of each market participant's notifications must match the final contract position. The utility notifies 60 MWh of production by Gen A, 40 MWh by Gen B, and 30 MWh consumption by Cons 1 and 20 MWh by Cons 2. The balance of the utility's notification is net production of 50 MWh, which offsets its final contractual position. The retail supplier schedules 50 MWh consumption by Cons 3, matching its contract balance, and the trader does not notify any physical position.²²

The net physical position of each player matches its contract position. Therefore, just as contract positions are balanced over the system, total production and consumption over the system nominated at gate closure are also balanced. This is illustrated in Figure 2.10.

The production and consumption programmes notified after gate closure serve two important purposes. First, they represent each market participant's commitment to deliver the electricity sold on the market and to consume the electricity purchased on the market. In the next section we discuss how those commitments are enforced.

Second, the notified production and consumption programmes tell the system operator where in the network market participants intend to

Figure 2.10 System net physical position at gate closure

produce and consume electricity at the time of delivery. This information is crucial to assessing whether the market participants' intended injections and withdrawals violate any system security constraints. In Section 2.3.4 and in Chapter 4 we discuss the remedial actions implemented by the system operator in such a case.

2.3 SYSTEM OPERATIONS

In this section we describe the activities performed by the system operator to ensure that production and consumption are in balance at all times. The three groups of system operation activities discussed in this section are illustrated on a time scale in Figure 2.11. They are imbalance settlement, balancing, and operating reserve procurement.

The objective of imbalance settlement is to identify any differences between each market participant's production (or consumption) commitments and actual production (or consumption), and to charge for them a price reflecting the value of electricity at the time of delivery. The objective of the balancing activity is to procure the energy needed to offset imbalances in real time. The objective of reserve procurement is to ensure that sufficient capacity will be available in real time in order to perform balancing. Both before and after gate closure the system operator carries out congestion management actions in order to ensure that that network's security constraints are met.

Figure 2.11 The timeline of energy delivery and ancillary services

We discuss system operations activities in reverse chronological order, starting from imbalance settlement (Section 2.3.1), moving on to balancing activities (Section 2.3.2), and finally the forward procurement of operating reserves before gate closure (Section 2.3.3). The discussion of the system operation activities in the first three sections ignores network security constraints. In Section 2.3.4 we discuss how system operation activities are impacted by network security constraints, and in Chapter 4 we analyse in greater detail the economic implications of network security constraints.

2.3.1 Imbalance Settlement

Production and consumption programmes notified after gate closure represent each market participant's commitment to deliver the electricity sold on the market and to consume the electricity purchased on the market. However, there is no guarantee that at the moment of delivery the exact nominated volumes will be produced and consumed and that there will be a perfect match between total production and total consumption in the system.

Departures of actual consumption and production from the notified volumes can and do happen after gate closure. For example, such imbalances could happen because of unforeseen generating unit outages, unexpected changes in weather conditions affecting units' performance,²³ and the availability of intermittent renewable energy sources such as wind and sunlight. On the consumers' side, imbalances reflect varying electricity requirements.

For most goods, a consumer whose supplier fails to deliver the contracted quantity can choose either to give up consumption of the missing quantity or to procure it from an alternative supplier. In the electricity markets, the latter option is implemented. However, because of the very short timing of real-time operations, such alternative procurement is centralised. In the event that the delivered or consumed quantities do not match the contracted quantities, the system operator is in charge of

procuring the replacement energy or disposing of the excess energy. This activity entails a series of transactions in which the system operator is the counterparty to the market participants.

The imbalance settlement process consists of two steps: assessment of each market participant's imbalance volumes and setting of the prices charged for the imbalance volumes. Below we present the main elements of imbalance settlement systems. We conclude this section by presenting the arrangements, known as load profiling, conventionally implemented to determine the smaller consumers' hourly electricity consumption for the purpose of assessing their suppliers' position on the wholesale market.

Imbalance volumes

After energy is produced and consumed, the system operator assesses the energy imbalance volumes for each market participant as the difference between its net nominated volume and its actual net energy production or consumption. Nominated volume is the net quantity of electricity that a market participant is committed to produce or consume as a consequence of its sales and purchases on the market, as notified to the system operator at gate closure. Actual production and consumption are metered at the generator and consumer premises.

A market participant runs a positive real-time imbalance if it produces a quantity of electricity greater than its net sales on the market, or if it consumes less than its net purchases. Through the imbalance settlement process, excess production is sold to the system operator. Likewise, a participant running a negative real-time imbalance buys replacement energy from the system operator.

We shall use the example from the end of the previous section to illustrate the imbalance volume assessment. We assume, in addition, that the actual production and consumption of market participants in a given delivery hour is different from that nominated at gate closure and therefore different from each participant's net contract position. The utility's clients Cons 1 and Cons 2, respectively, consume 30 MWh and 25 MWh, compared with the nominated quantities 30 MWh and 20 MWh; the production of Generator A is 70 MWh instead of the nominated 60 MWh, and the production of Generator B is 10 MWh instead of the nominated 40 MWh. Consumption by the retail supplier's Cons 3 is 45 MWh instead of the nominated 50 MWh.

Figure 2.12 illustrates the calculation of the utility's imbalances in our example as the difference between net actual production and net nominated volume. The actual physical balance for the utility given by the sum of the production of the utility's generating units and of the consumption of the utility's clients is 25 MWh. The net contract position and net

Figure 2.12 Imbalance of the vertical utility

Figure 2.13 Imbalance of the retail supplier

nominated volume at gate closure was 50 MWh. The utility has therefore delivered 25 MWh less than it has sold in the market and notified. The system operator assesses the negative imbalance of 25 MWh, an amount that the system operator had to supply in real time in the utility's place.²⁴

Figure 2.13 illustrates a similar calculation of the imbalances for the retailer in our example. The retailer's customer has consumed 5 MWh less than the amount purchased by the retail supplier at gate closure. The retailer is then assessed as having a positive 5 MWh imbalance.

Each market participant's imbalance volume is known only after the fact, when metering information becomes available. Before that, in real time the system operator has offset the total system imbalance, that is, the sum of actual production and consumption of all participants, by

Figure 2.14 System-wide imbalance

Table 2.1 SO transactions in a simple imbalance settlement system

SO transactions	Volume, MWh	Price, €/MWh	Value, €
Purchase on the real-time market	-20	50	-1,000
Imbalance settlement: utility	25	50	1,250
Imbalance settlement: retail supplier	-5	50	-250
Total	0		0

procuring additional production or reduced production. As shown in Figure 2.14 for our example, the total system imbalance is the sum of all the market participants’ individual imbalances.

Imbalance prices

The prices at which imbalances are cleared with the system operator are called ‘imbalance prices’. Efficient imbalance prices reflect the value of electricity at the time of delivery. That value is the clearing price of the real-time market where the system operator buys and sells electricity in order to offset the total system imbalance.

Table 2.1 shows the outcome of transactions between the system operator and the market participants which run imbalances, and those carried out by the system operator to procure the energy to offset the total system imbalance. We have assumed that the real-time clearing price is €50/MWh. The sum of energy volumes bought and sold by the system operator is zero because it buys the volume of energy on the real-time market that is needed to match the total system imbalance, that is, the sum of the market

participants' individual imbalances. The system operator's budget is also balanced in our example, as all the transactions carried out to settle individual imbalances are priced identically.

In practice, assessing the value of electricity at the time of delivery for the purpose of setting imbalance prices is not as straightforward as in this example. One of the reasons is the temporal and spatial granularity mismatch between the imbalance prices and the prices of the transactions performed by the system operator on the real-time market. In the balancing market the system operator may have to buy or sell balancing energy at different prices in different intervals of the same hour²⁵ and at specific locations of the network,²⁶ whereas imbalance volumes are usually settled on an hourly or quarter-hourly basis, irrespective of where in the network the imbalance took place. Below we discuss how this mismatch is dealt with.

Dual imbalance pricing systems

Although trading between market participants stops at gate closure, they can buy or sell power after gate closure by running voluntary imbalances. For example, a generator may expect a large negative system imbalance in real time or, equally, that the clearing prices of the real-time market will be much higher than the day-ahead or intraday prices. It might then take the risk of producing more than the notified volumes. This would result in a positive imbalance by the generator that would be purchased by the system operator at a price reflecting real-time market prices, as illustrated in the previous section. Likewise, a generator expecting real-time prices lower than the day-ahead and intraday prices might decide to produce less than the notified volumes and to buy the energy deficit at the imbalance price.

In most European markets, with the exception of the Netherlands, such behaviour is discouraged through specific imbalance pricing arrangements. The main rationale for discouraging voluntary imbalances is the mismatch between the temporal and locational granularity of the prices in the real-time balancing market and the prices charged for the imbalances. Because of this mismatch, the price charged for an imbalance may not correctly reflect the cost, or reduction in cost, effectively caused by the imbalance in real time. This may therefore create distorted incentives to market participants to run voluntary imbalances, especially in the event of congestion in the network, as we discuss further in Section 2.3.4 and in Chapter 4.

Market participants are encouraged not to run voluntary imbalances by specific imbalance-pricing arrangements that make imbalances less profitable than sales and purchases on the market, irrespective of the cost or cost saving that these effectively bring about. One such mechanism sets a different price for imbalances of a given participant depending on the positive/negative system imbalance. In the industry jargon these arrangements

Table 2.2 Dual-price imbalance settlement system

		System imbalance	
		Negative ($P_{BM} > P_{DA}$)	Positive ($P_{BM} < P_{DA}$)
Market participant imbalance	Negative	P_{BM}	P_{DA}
	Positive	P_{DA}	P_{BM}

are commonly referred to as ‘dual imbalance pricing systems’. A stylised example of this approach which is implemented in many European countries is shown in Table 2.2. The imbalance charges presented in this table are based on the combination of the price of the day-ahead market P_{DA} and the price on the real-time balancing market P_{BM} .

When the system imbalance is negative, the system operator purchases additional power on the real-time balancing market to balance energy in the system. The real-time price on the balancing market in this case is likely to be higher than the day-ahead price. Market participants that run negative imbalances pay the balancing market price, which is higher than the price they would have paid on the day-ahead market to purchase the electricity that they did not deliver (or that they consumed in excess of their purchases). Market participants running positive imbalances and reducing the system imbalance receive the day-ahead market price and do not profit from their behaviour. They would have obtained the same profit by selling the additional volumes on the day-ahead market.

When the system’s net imbalance is positive, the opposite applies: market participants responsible for positive imbalances receive a balancing market price that is lower than the day-ahead market price for production in excess of notification, while market participants running negative imbalances pay the day-ahead market price and thus do not benefit from helping to balance the positive system imbalance.

In this model the imbalance prices are explicitly designed to eliminate any possibility for market participants to profit from a voluntary imbalance in any direction.

Examples of such systems can be found in France, Belgium and Italy. The dual cash-out system that is used in the UK is similar. A participant with a negative imbalance, when the system imbalance is also negative, pays the System Buy Price, determined by the price of accepted offers on the real-time balancing market. This price is typically higher than the forward market price. A market participant with a positive imbalance, when the system imbalance is also positive, receives the System Sell Price based on the accepted bids on the real-time balancing market. This price is normally

Table 2.3 *SO transactions in a dual-price imbalance settlement*

SO transactions	Volume, MWh	Price, €/MWh	Value, €
Purchase on the balancing market	-20	50	-1,000
Charge to the vertical utility	25	50	1,250
Payment to the retail supplier	-5	30	-150
Total	0		100

lower than the forward market price.²⁷ Finally, when the participant's imbalance is the opposite of the system imbalance, the participant pays or is paid the reverse price, which is a market index price based on short-term energy trades on the within-day spot markets, like the day-ahead price in the system described above.

Consider the example above, assuming a dual-price imbalance settlement. In our example, the system imbalance is negative. We assume that the price in the balancing market is €50/MWh and the price in the day-ahead market is €30/MWh. Table 2.3 summarises the value of the transactions carried out by the system operator on the balancing market and in the imbalance settlement. In this case, the utility pays a high real-time imbalance price for the negative imbalance, while the supplier gets paid a lower day-ahead price for its positive imbalance. The table shows that in the case of the dual-price imbalance mechanism, the system operator budget no longer automatically balances and shows a surplus. Such surpluses are generally passed on to the network users, for example in the form of lower transmission tariffs.

A theoretical study carried out by Vandezande et al. (2010)²⁸ analyse the main impacts of dual pricing on wholesale trade. In addition to the lack of balance in the system operator's budget, the study mentions possible discrimination in favour of larger players in the event that their imbalance volume is assessed at portfolio level. Larger market players sustain lower imbalance costs because they have more opportunity to self-balance within their portfolio. The dual-price schemes provide incentives for inefficient strategies such as overcontracting in the wholesale market, withholding services for own use and nominating less than the expected production.

Portfolio aggregation for imbalance settlement

When different prices apply to negative and positive imbalances, market participants face different imbalance costs, depending on the degree of aggregation of production and consumption for the purpose of imbalance volume assessment.

In the example above, the imbalance volume was calculated at the

Table 2.4 Imbalance settlement of utility on a plant and customer level

Energy account	Notified volume, MWh	Actual volume, MWh	Imbalance volume, MWh	Price, €/MWh	Imbalance Cost, €
Generator A	60	70	10	30	300
Generator B	40	10	–30	50	–1,500
Customer 1	30	30	0	–	–
Customer 2	20	25	–5	50	–250
Total			–25		–1,450

market participant's portfolio level, that is, as the sum of the imbalances of all generating units and consumers under the market participant's responsibility. Under portfolio-based imbalance volume assessment, how production is split between Gen A or Gen B is irrelevant to the utility. That is because, given the consumption of the utility's customers, only the sum of the two units' production is relevant in assessing the imbalance volume.

However, reallocation of production between generating units in real time may create additional system operation costs, especially if it leads to violations of network security constraints. The system operator may then want to discourage voluntary imbalances, not only at the aggregate portfolio level, but also at the level of individual generating units and consumers. For this reason, in some markets the system operator assesses the imbalance volume at a finer level of granularity.

For example, the imbalance volume can be assessed separately for the portfolio of generators and for the portfolio of consumers under each market participant's responsibility. Alternatively, in some markets the imbalance volumes of large generating units are assessed individually, while the net imbalance is assessed for large sets of consumers.

Consider the vertical utility in our example above. At portfolio level it runs a negative imbalance of 25 MWh. In the case of dual imbalance prices in the example above, this imbalance will be charged at a price of €50/MWh, and the total cost imbalance for the utility will be €1,250. However, the negative imbalance of 25 MWh is the sum of positive and negative imbalances at the level of individual plants and customers, determined as the difference between notified volumes and the volumes achieved in real time for each generator and customer. Table 2.4 summarises the imbalance settlement for the utility performed at the level of individual generators and customers.

Generator A creates a positive imbalance that is charged at a lower price of €30/MWh, whereas the negative imbalance induced by Generator B

and Consumer 2 is charged at a higher imbalance price of €50/MWh. As a result, the total imbalance cost of the utility in this case is higher than if it were assessed at portfolio level. Even if there was a perfect energy balance within the utility's portfolio, the entity would still face imbalance costs if it distributed production between generating units differently from the way it nominated production at gate closure.

European markets that assess producers' and consumers' imbalances separately include Spain, Italy and the Nordic countries. In Italy, imbalances are assessed at the generating unit level for large generators and at plant level for large customers. On the contrary, for the purposes of assessing imbalances, injections by smaller units and withdrawals by small consumers are netted out. In the UK, individual generator energy accounts are calculated for each power plant. In France, Germany, the Netherlands, Austria and Belgium, imbalances are assessed on the total portfolio, which combines generators and consumers. In addition, multiple market participants can form a balance responsible party, a legal entity taking financial responsibility for the imbalances of consumers and generators within its perimeter.

Load profiling

Assessing the imbalance volume of a market participant involves comparing the notified volumes with the actual production and consumption volumes during each hour or smaller time interval. We have referred to this interval as the 'balancing period', and have assumed hourly balancing periods to simplify the description.

The actual output of most power plants is metered continuously. Therefore assessing the actual production of generators in each 'balancing period' presents few technical issues. Assessing consumer imbalances is less straightforward. Consumption by small electricity customers is rarely metered on an hourly basis. Conventional mechanical meters that are still widely in use can only record the total energy consumption between two meter readings that are typically taken on a monthly or multi-monthly basis.²⁹

However, being able to assess the hourly consumption of each end-user is a necessary condition for implementing retail competition. It would be impossible otherwise to assess whether a volume of electricity matching the consumer's withdrawal had been procured on the wholesale market by the consumer's supplier. Since the value of electricity on the wholesale market is different in each balancing period, the withdrawal for which the retailer is responsible, that is, its clients' electricity consumption, must be assessed for each and every balancing period.

For this reason, arrangements referred to as 'load profiling' have been

Table 2.5 Load profiling of a non-hourly metered customer, MWh

	Past consumption	Hour 1	Hour 2	Hour 3	Hour 4
All non-hourly metered customers	5,000	1,000	1,200	1,400	800
Customer A	100				
Share of consumption allocated to Customer A	2%				
Hourly consumption allocated to Customer A		20	24	28	16

developed in order to calculate and settle consumption by non-hourly metered consumers. In simple terms, for each hour each consumer is allocated a share of the withdrawals performed by the entire set of non-hourly metered customers. The consumer’s share of the total consumption is determined on the basis of its consumption history. The total non-hourly meter consumption for each hour is calculated as the difference between total production and total hourly metered consumption.³⁰

Table 2.5 illustrates this process in an example where, for reasons of simplicity, we have assumed that the total consumption of each non-hourly metered costumer is measured not every month, but every four hours. The table shows that the past metered consumption of Customer A represented 2 per cent of the past consumption of all non-hourly metered customers. Therefore the customer has been allocated 2 per cent of consumption by all non-hourly metered customers in each of the following four hours.

The retail supplier serving Customer A will be considered to have an imbalance equivalent, for each hour, to the difference between the nominated consumption and the consumption allocated to the consumer by load profiling.

At the end of the fourth hour the total consumption by Customer A in the four hours is known. The metered consumption will generally be different from the consumption allocated by the profiling system, as the latter was based on the consumer’s past share of total non-hourly metered consumption. Table 2.6 illustrates the approximation error that may occur as a result of load profiling. In this example the total consumption over four hours approximated by load profiling for Customer A is 88 MWh, while the actual metered volume of consumption over those four hours was 120 MWh. The 32 MWh difference represents the volume that Customer A has consumed in addition to what its retail supplier has notified on its customer’s behalf.

Table 2.6 Volume reallocation, MWh

	Hour 1	Hour 2	Hour 3	Hour 4	Estimated volume	Actual volume	Diff.
All non-real-time metered customers	1,000	1,200	1,400	800	4,400	4,400	0
Customer A	20	24	28	16	88	120	32

Note, however, that the total consumption allocated by the load-profiling scheme is, by construction, equivalent to the actual consumption of the entire group of non-hourly metered consumers. That means that some other non-hourly metered customers have consumed 32 MWh less than they were allocated by the profiling system. These differences are then settled among the retailers of the non-hourly metered consumers, by applying a volume-weighted average day-ahead price. In our example the retailer serving consumer A will pay for the 32 MWh consumed by its client (but not allocated to it by the profiling system), while some other retailers will be paid for 32 MWh allocated to but not consumed by their clients. This process of settling the errors resulting from load profiling between retail suppliers is sometimes referred to as ‘volume reallocation’.

We note incidentally that load profiling is not a way to address the implications of the lack of demand price sensitivity discussed in Section 2.2.1. The cost to a retailer of serving a load-profiled consumer does not depend on how the consumer’s actual consumption is spread across the hours. The supplier is financially responsible to the system operator only for the time profile conventionally assigned to the consumer and not for the actual pattern of the consumer’s withdrawals, which remains unknown. As a consequence, the supplier does not benefit from its load-profiled consumer shifting consumption from high- to low-price hours.³¹ The same holds for the consumer, whose supply price cannot be made contingent on the actual pattern of consumption, which is not measured, but only on the total consumption for the month.

2.3.2 Real-time Balancing

As discussed above, even if volumes bought and sold in all market transactions between participants are balanced, production may not perfectly match consumption at the time of delivery. In this section we describe how the system operator procures energy to balance production with

consumption in real time to ensure that the power system is secure at all times. We present first the different services supplied by the generators to the system operator in real time. Then we discuss the market arrangements through which those services are procured.

Balancing services

Some balancing happens automatically with little direct involvement of the system operator. This type of balancing is called ‘automatic control’. Automatic control is driven by changes in some system physical conditions caused by the energy imbalance. Automatic control acts fast, providing the time for the system operators to give instructions for manual corrections, after which injections by units providing automatic control are restored to their scheduled levels, ready to respond to a new imbalance.

In Europe a distinction is often made between two types of automatic control: primary and secondary control.

Primary control is an automatic reaction to frequency deviation of some generating units that are already producing. Imbalance between production and consumption causes the system’s frequency to deviate from the nominal value (for example, 50 Hz in Europe). Frequency tends to fall in the event of negative imbalance and to increase when the imbalance is positive. Units providing primary control automatically increase production when system frequency falls and decrease production when frequency increases, offsetting the imbalance that has caused the frequency change. Primary control is normally activated within 30 seconds following the disturbance. A generator providing primary control is expected to perform deviations from its production programme of very short duration, around 15 minutes, after which primary control is replaced by other types of control.

A frequency deviation causes all units capable of providing primary control to respond, irrespective of where the imbalance between production and consumption has occurred in the interconnected network. As a consequence, the automatic activation of primary control may change the power flows between the country that experienced the imbalance and the neighbouring countries, resulting in the unintentional import and export of electricity.

Secondary control is a system that remotely controls specially equipped generating units located in a control area, in order to regulate their output up or down depending on the interchange flows between the control area and the neighbouring areas. For example, in response to a sudden deficit in production in a control area, primary control would increase production by all the generating units connected to the network within and beyond the control area. This would increase the import flows into the area where the production deficit occurred compared with the levels of

Figure 2.15 Sequence of automatic and manual control activation

cross-border transactions notified to the system operators by the market participants. Secondary control adjusts generation in the control area where the deficit has occurred until the cross-border flows are restored to their scheduled values. Generators supplying secondary reserve must be able to vary output within 15 minutes of the signal being sent and maintain the new production level for several minutes.

While automatic control is providing nearly instant response to the energy imbalance, the system operator gives instructions for manual changes to the production levels of the available units in order to offset the initial imbalance. In Europe, such manual control is often called ‘tertiary control’. After manual corrections are activated, automatic primary and secondary control are no longer needed and are restored to their original levels in order to be ready for activation in the event of another contingency.

Figure 2.15 illustrates the typical sequence of activation of automatic and manual control in the event of a real-time imbalance, such as the outage of a large generating unit. The generator’s outage causes a sudden drop in the system’s frequency that triggers the activation of primary control. Within minutes the secondary reserve capacity located in the country where the outage has occurred increases output. As secondary control increases production, the frequency rises back towards the target level and the country’s net imports return to their scheduled levels. Once the frequency reaches the target level, the production by the units providing primary control goes back to the target schedule and the primary control reserve margin is fully restored, ready for activation in case there is further need. Following the outage the system operator has also provided manual instructions to a slower generator to increase production. As soon as this tertiary energy control is implemented, production by the generators providing secondary reserve automatically reduces until

their initial schedules have been fully restored. At that point production by the missing unit has been entirely replaced by the manually activated generators.

The balancing market

The energy produced by the primary reserve capacity is commonly not paid for, under the assumption that production increments and decrements, compared with the scheduled programmes, will balance out. Typically the electricity delivered by units providing secondary control is paid for either at a predetermined price or at the same price as the energy delivered by the generators providing tertiary reserve.

Manually controlled energy output increases and decreases are procured by the system operator on the balancing market, where generators, and possibly consumers, submit offers to provide upward and downward regulation.³² The upward offers are prices that the generator agrees to receive for an increase in production from a specific unit compared with the energy schedule nominated at gate closure. Conversely, downward bids are the prices that the generator is willing to pay in order to decrease production from a specific unit compared with the nominated schedule.

If the system imbalance in the control area is negative, that is consumption exceeds scheduled production within the control area, the system operator eliminates the imbalance by accepting the necessary amount of cheapest upward regulating bids. Similarly, in the event of a positive overall imbalance in the control area, the system operator accepts downward regulating bids. Since accepting a downward regulating bid amounts to a sale of electricity to the bidder by the system operator, the latter selects the highest-priced downward regulating bids, in order to reduce the output of the most expensive units. This allows the system operator to keep the system balanced at minimum cost.³³

As in the case of the day-ahead markets, there are two main options for settling accepted regulating bids and offers: the pay-as-bid and single-price clearing mechanisms. We looked at the relative merits of both approaches in the context of day-ahead markets in Section 2.2.1. The single clearing-price mechanism is generally regarded as more efficient for implementation in short-term electricity markets. However, for the balancing market many European countries have opted for the pay-as-bid system (for example, Italy, the UK, France and Belgium). The preference for the pay-as-bid system in the context of the balancing markets appears to be based on some specific features of the system operator's demand for balancing services, such as the need to accept bids and offers on a continuous basis and to address network security constraints, in addition to ensuring energy balancing, as described further in Section 2.3.4 and in Chapter 4.

The alternative design of the balancing market implemented in the United States is based on a series of non-discriminatory auctions, typically run every five minutes. Each session computes the system marginal price at each network location. That price is paid for all the offers accepted in that session and charged for all the bids accepted in that session. The same price is also applied to all imbalances. We discuss this approach in greater detail in Section 2.4.

Market arbitrage between the real-time and wholesale markets

Since electricity is non-storable, the price prevailing in the balancing market at each time is, strictly speaking, the only spot price for electricity. All transactions, with the exception of those in the balancing market, have a forward nature, given that electricity cannot be transferred from the seller to the buyer before real time, that is, when it is consumed.

Despite the fact that quantities traded in the balancing markets are generally small, the prevailing balancing prices, or real-time prices, may have a strong impact on prices in the wholesale electricity markets. Market participants can buy and sell electricity at real-time prices by making a bid or an offer on the real-time market, or by voluntarily running imbalances in real time and facing imbalance prices that are to some extent determined by the real-time prices (see Sections 2.3.1 and 2.3.4).³⁴ In other words, market participants have the opportunity to arbitrage between the wholesale and the real-time markets. No generator would want to sell on the wholesale market at a price lower than the expected real-time price, and no consumer would want to buy on the wholesale market at a price higher than the expected real-time price. As a consequence, any distortions in the real-time prices may filter through to the wholesale electricity prices.

2.3.3 Procurement of Operating Reserve Capacity

Real-time balancing services are provided by flexible generating resources that can quickly change output within the timeframe of the balancing market.³⁵ Flexible generators are, for example, hydropower plants and combustion turbines that can be brought into service at very short notice. Balancing services can also be provided by slow thermal generators that are already scheduled to operate below their maximum operating limit and can change their output within the required timeframe.

In order to ensure that in real time there is sufficient unloaded capacity to perform balancing, the system operators procure the reserve capacity long before gate closure. Generators providing such operating reserves undertake not to sell part of their capacity on the market and to keep part-loaded capacity available for the system operator's balancing purposes.

Below we discuss types of operating reserves as well as the cost of providing reserves for generators. We also discuss the existing approaches used by system operators to procure these reserves.

Types of operating reserves

Types of operating reserves differ in the dynamic characteristics of the reserved generating capacity. Generating capacity reserved to provide primary, secondary and tertiary control must be able to ramp up or down to the amount of reserve quantity provided, within the timeframe of the type of control that needs to be provided in real time, for example, 30 seconds in the case of primary reserves and 15 minutes in the case of secondary reserves.³⁶

Therefore, the amount of each type of reserve that a given plant can provide is limited by its ramping rates. Because of the limit on the ramping rates, a single unit can rarely meet the entire reserve capacity requirement, and the system operator needs to procure reserve capacity from many units.

Resources supplying primary and secondary reserve capacity undertake to make the reserved capacity available in real time and to respond to the automatic control activation either upwards or downwards. Resources supplying tertiary reserve capacity are required to offer the corresponding production volume on the balancing market.

In most implementations the reserve capacity commitment does not fix the price for the electricity produced in the event of activation, that is, the price indicated in the balancing offers. However, the selection of reserve capacity offers may also take into account the price of electricity delivered in the event of activation. In this case, the price paid to generators to provide balancing energy, if activated in the real time, is also set by the reserve market³⁷

Cost of providing reserve capacity

The cost of providing primary, secondary or tertiary operating reserves for generators is mainly determined by the opportunity cost of keeping capacity part loaded. In the case of reserve capacity for upward regulation, a generator needs to maintain the energy schedule below its maximum capacity, in other words providing ‘capacity headroom’³⁸.

For example, if the spot-market price is €50/MWh, a generator with a variable cost of €30/MWh could receive a profit of €20/MWh by selling its power on the market. This means that providing 1 MW of upward operating reserve capacity per hour would entail opportunity cost of €20/MWh per hour. This reserve provision opportunity cost arises when the wholesale market is profitable for the generator, that is, when the generator’s variable cost is below the energy spot price. In the opposite case, when the

energy market price is lower than plant variable costs, providing operating reserves may entail the cost of starting the unit up and the net cost of selling the minimum load at a price below the variable cost of production.

Opportunity costs are zero when the reserve capacity is provided by marginal and out-of-merit units, that is, by generating units with a variable production cost equivalent to or higher than the energy market-clearing price. However, the ramping rates limit the amount of reserve capacity provided by each individual unit. As a result, the requirement for reserve capacity cannot always be met by marginal units alone. Part of the reserve capacity is thus met by part-loading units with a high opportunity cost. For those units, the opportunity cost determines the cost of providing operating reserves.

Procurement of operating reserves

European system operators use two main ways to procure reserves. One is to buy reserves before the day-ahead market by acquiring unloaded capacity from market participants. After the reserves have been acquired it becomes the responsibility of the market participant to ensure that after clearing of the day-ahead and intraday markets the contracted reserve capacity is available. It is also up to the market participant to decide which units to use to schedule the reserve. Reserves can be contracted from years to hours before the time of delivery.

The other approach used by the Italian, Irish and Spanish system operators is to build the reserve margin after clearing of the day-ahead market when preliminary day-ahead plant schedules become known. After closure of the day-ahead market the system operators in these countries run special markets where they buy generators' deviations from the preliminary day-ahead schedules. For example, by purchasing a decrease of scheduled production by a unit, the system operator creates capacity headroom which will serve as reserve. These markets are called the 'restriction market' in Spain and the '*ex ante* MSD market' in Italy. They are also used to perform system re-dispatch before gate closure as we discuss below in Section 2.3.4.

The close interaction between the production of energy and provision of operating reserves provide the rationale for market arrangements for procuring these two products to be closely integrated. Integrated markets ensure the efficient allocation of generating capacity between the production of energy and operating reserves. If markets for operating reserves and energy markets are instead cleared independently, generators have to face risks, because assessing the profit-maximising offer price for one product means forecasting the price of the other. Forecast errors may then lead to inefficient offer and production decisions.

Perfect integration between energy and operating reserve markets is achieved on some of the US markets, where the spot energy market is cleared simultaneously with the reserves market. The system operator, which in the US markets also performs the function of market operator, selects the generators' offers for power production in order to optimally allocate generation to meet several objectives simultaneously: meeting the energy demand as well as the requirement for each type of reserve capacity. As a result of this simultaneous clearing, the system operator identifies the energy schedule and the operating reserves capacity schedule for each plant. It also identifies the market-clearing prices for energy and for each of the operating reserves.

In the European markets, the energy markets and the markets where the system operator procures operating reserves are typically cleared in sequence. The inefficiency of this approach is mitigated by bringing the market for operating reserves as close in time to the spot market as possible. This makes it easier for the market participants to arbitrate between the energy and the reserve markets, by increasing the quality of the predictions of the clearing prices of the two markets. This is done, for example, in Germany, where tertiary reserves are procured in a day-ahead timeframe, shortly before clearing of the day-ahead energy market (minute reserve). Generators submit offers to make capacity available in real time. The offers specify the quantity offered and the unit price. The system operator accepts the set of offers that meet the target levels of the different types of reserves at minimum cost. The approach implemented in Italy and Spain, where the system operator runs special markets shortly after the day-ahead market, also makes it possible to reduce the impact of the segregation of energy and reserve markets.

However, in many other European countries, operating reserves are procured by the system operator under long-term contracts. For example, in France primary and secondary operating reserves are procured under three-year contracts, and in Belgium, primary and tertiary reserves are contracted over four years, and secondary reserves over two years.

Other ancillary services

In addition to operating reserves, ancillary services also comprise the provision of reactive power and black-start capability.

The provision of reactive power is necessary to correct the voltage level on transmission lines. As the power flows from generators to consumers the voltage on the grid tends to drop. If the voltage drop is not corrected it may result in unsatisfactory operation of electrical equipment, causing damage to electrical motors. The voltage drop on a transmission line can be corrected by transformers that automatically adjust the voltage before

Figure 2.16 Example of re-dispatch

the power is distributed and through the injection of reactive power by generators. Reactive power is generally produced at a relatively low cost.

Black-start capability services are needed in the event of black-out to provide power to the grid in order to allow other plants to restart producing power. Most thermal generators cannot start up without taking electricity from the grid. Low-cost black-start capability is normally offered by hydropower plants.

2.3.4 Network Security Constraints and Market Design

Throughout the chapter we have mentioned that, in addition to matching production and consumption, the system operator guarantees the security of the network. When one or more network constraints are binding, the cost of meeting an increase in demand can be different at different nodes, as the example in Figure 2.16 shows. In this example, there are four power plants with capacity of 200 MW, each located in two interconnected zones. The variable costs of the two generators in zone A are, respectively, €10/MWh and €30/MWh, and the variable costs of the generators in zone B are €20/MWh and €50/MWh. Demand in zone A is 50 MWh and in zone B is 450 MWh. If there were no transmission capacity constraint between the two zones, the overall demand of 500 MW would be met at the least cost by generators A1, B1 and A2. If these generators bid their variable cost in the day-ahead market, the cost of generator A2 would set the day-ahead price at €30/MWh.

In this case, production in zone A would be 300 MW and production in zone B would be 200 MW. Given the consumption in each zone, this production plan would imply an export of 250 MW from zone A to zone

B. This export might not be feasible in the event that the network capacity between the two zones was 100 MW. Given this network security constraint, the least-cost option to meet the demand would be to produce 150 MWh in zone A and 350 MWh in zone B. In this case the marginal unit of energy produced in zone A would cost €10/MWh, that is, less than the clearing price that would be applied in the absence of the constraint, and the marginal cost of energy produced in zone B would be €50/MWh, that is, greater than the unconstrained clearing price.

In Chapter 4 we provide a more general description of the relationship between network constraints and the incremental cost of meeting demand at the different network nodes.

Alternative market designs differ in how they handle network constraints. The analyses carried out in this chapter thus far are consistent with an approach extensively implemented in Europe, in which the electricity market is run ‘as if’ no network constraints are binding³⁹. In other words, market transactions are not subject to any network-related restrictions. This is achieved by standardising the traded product across locations. Market players buy and sell electricity knowing that the delivery and withdrawal obligations corresponding to their sales and purchases can be fulfilled at any location of the network of their choice. As a consequence, the energy market clears with a single system-wide price, as injections at any location are assumed to be perfect substitutes, as are withdrawals.

In this section we discuss the arrangements implemented within this model to ensure that system security conditions are met at all times. First we analyse how network congestion affects system operations, and then we go on to discuss how they affect imbalance price calculation.

System re-dispatch

When market transactions are not subject to any network-related restrictions, production and consumption schedules that market participants notify to the system operator at gate closure may be infeasible, that is, they may result in power flows that violate some transmission constraints. This situation is known as ‘congestion’. Congestion is relieved by reallocating production between generators located at different nodes of the network, in order to ensure that the resulting power flows are within the network capacity limits.

This is achieved by paying generators⁴⁰ to modify their production from the levels notified at gate closure. These transactions result in different prices at different network nodes, because the congestion can only be relieved by increasing production at certain nodes and reducing production at certain other nodes.⁴¹

Such re-allocation is referred to as ‘system re-dispatch’. System operators

perform the system re-dispatch both after gate closure, during real-time balancing, and through various arrangements before gate closure.

In the real-time balancing market the system operator selects bids and offers to depart from the nominated volumes submitted by various plants in order to keep the system secure at minimum cost. In the previous section we assumed that the overall energy balance was the only relevant security constraint. In reality, in the real-time balancing market the system operator also seeks to meet the network security constraints. When one or more of such constraints appears to be violated, the system operator accepts bids and offers at selected network nodes until the constraint is relieved. In that case the system operator needs to pay location-dependent prices, acknowledging that production at different locations is not interchangeable for the purpose of relieving a given set of network constraints. For example, a high-price offer may have to be accepted by the system operator simply because it is for a unit located in the same area that an imbalance occurred, while an identical imbalance at a different location could be offset by a much cheaper unit.

Locational pricing in the real-time market can be achieved in two ways: determining a location-specific clearing price that would be paid to or received from all accepted units in the same location, or clearing each accepted offer or bid at the offered or bid price, that is, using the pay-as-bid method. The former approach is implemented in the US markets, whereas the latter is used in most European balancing markets. As a result, the prices of the accepted balancing bids and offers may vary significantly across network locations.

Some network issues and constraints can be predicted by the system operator long before gate closure, for example, based on preliminary day-ahead production schedules obtained after the day-ahead market is cleared. It may be cheaper to address such congestion before real time by changing the commitments of cheap but slow units before gate closure, rather than performing re-dispatch of fast but expensive units in real time.

System re-dispatch applied before gate closure is used in several European markets, but approaches differ widely from country to country. Italy and Spain use a rather structured approach where, immediately after clearing of the day-ahead markets and after obtaining day-ahead generation schedules, the system operator runs an organised market for constraints resolution, accepting bids and offers from plants at specific locations in order to resolve network congestion. These markets are the *ex ante* MSD market in Italy and the restrictions market in Spain, and are the same markets where the system operator in those countries procures operating reserves.

Other countries use less-structured approaches. For example, in the UK

the system operator may conclude a contract to acquire output from a particular plant before gate closure in order to resolve expected transmission issues. Such contracts are called ‘Pre-Gate Closure Balancing Trades’, or PGBTs.

The need for re-dispatch actions prior to market gate closure increases as congestion is expected to increase because of changes in the generating fleet, such as the deployment of a large amount of intermittent renewable power or phase-out of nuclear power not immediately matched by the necessary network upgrades. In several European countries, such as Germany, the Netherlands and Switzerland, system operators are currently considering introducing arrangements for constraint resolution before gate closure.

Early re-dispatch actions performed before gate closure may be followed by other trades between market participants in the intraday timeframe. Unless there are any specific restrictions, such trades may result in final nominations that restore the pre-re-dispatch flows, undoing the effect of early re-dispatch actions. Therefore, an effective early re-dispatch may require a mechanism enforcing the schedules achieved through re-dispatch throughout the intraday timeframe without unduly restricting the intraday trades. For example, one way to achieve this is to allow the system operator not to accept changes in final unit notifications compared with the preliminary day-ahead notification if such changes undo the effects of re-dispatch actions.

Imbalance prices in re-dispatch-based systems

In most European countries, market transactions are performed based on the assumption of unlimited transmission capacity, and are therefore not subject to any network-related restrictions. As a result, a single price clears electricity demand and supply in the entire country. Any re-dispatch cost incurred by the system operator in the event that market participants’ intended production and consumption violate some network constraints is then socialised among all the network users.

The convention that location has no impact on the value of electricity needs to be carried over to the imbalance settlement stage, in order to avoid distortions in the market participants’ behaviour and to limit the cost of balancing for the system operator. Consider, for example, the case in which imbalance prices are set at a national level in order to reflect the real-time balancing prices as described in Section 2.3.1 above. In the event of congestion, such imbalance prices would not be fully cost reflective. In the example in Figure 2.16 this means that the imbalance price charged by the system operator to generators for the negative imbalance in zone B would be lower than the price to increase production in this zone in order to resolve the imbalance. Likewise, the imbalance price paid by the system

operator to generators for the positive imbalance in zone A would probably be higher than the balancing costs in this zone.

Imbalance prices set at a national level cannot therefore be relied upon to send correct signals to market participants of the cost caused by their imbalances in the event of congestion. In fact, making imbalance prices cost reflective would mean allowing them to vary by location. However, if energy markets are still cleared at uniform national prices, this would produce arbitrage incentives with potentially adverse wealth transfer from consumers to generators, as we discuss in detail in Chapter 4 Section 4.4. In addition, allowing arbitrage through voluntary imbalances could make system operations more complicated, because notifications would not provide reliable information about actual production at the time of delivery.

For that reason, in markets where congestion is dealt with via re-dispatch, imbalance pricing mechanisms like the dual pricing system discussed in Section 2.3.1 are implemented in order to discourage arbitrage via voluntary imbalances.

Once locational cost-reflective imbalance charges are ruled out, assessing the imbalance price entails some level of discretion. In the UK, for example, the system operator labels each transaction in the real-time market as either system or energy balancing. System-balancing transactions are those related to the resolution of network constraints, and are therefore removed from the calculation of imbalance prices.⁴² Note, however, that the labelling system is highly conventional, since it is conceptually impossible to attribute a single purpose to real-time transactions that are chosen to address multiple simultaneous constraints.

2.4 WHOLESALE MARKET DESIGN IN EUROPE AND THE US

Throughout this chapter we have highlighted differences between the European and the US wholesale electricity market design. In this section we provide a brief analysis of the Standard Market Design, developed by the Federal Energy Regulatory Commission (the FERC) in 2003. Although this proposal was not officially implemented, many of the electricity markets currently operating in the United States follow the principles of the Standard Market Design.

The European and the US market designs differ in how they address the special technical features of electricity. The European design draws a clear line between the day-ahead and intraday energy markets on the one hand and the ancillary services and balancing markets on the other. The design

of the day-ahead and of the intraday markets emphasises that electricity is a commodity; the special features of electricity are catered for by the ancillary services and real-time markets. In the US approach, on the contrary, the markets running from the day-ahead market through to real time are highly integrated. They share the same auction format, clearing algorithm and network security constraints. More importantly, they share the same product definition, which is fully consistent with the technical features of electricity. As a result, the specific technical features of electricity are consistently addressed in all market venues.

We compare next some features of the US and the European approaches as far as market transactions are concerned, to the extent that some common features can be identified in the very diverse arrangements implemented in Europe.

2.4.1 Addressing Production Constraints

European energy markets are largely based on bilateral transactions and voluntary organised exchanges that trade standardised products, such as hourly bids and sometimes block bids. Central to US energy markets is a pool where generators submit the bids that are closely related to the individual generating plants, specifying the detailed plant dynamic characteristics in addition to the bid-based costs. The market-clearing algorithm takes into account the detailed constraints of the production fleet and looks for a solution that meets demand at the lowest cost while ensuring that production programmes are feasible.

2.4.2 Energy and Balancing Markets Integrated with Network Constraints

The US and the European approaches differ in how network security constraints impact the market outcome. The European energy markets impose no limits on trading within the entire market area, even in the event that such transactions result in power flows that violate network security constraints. Such network security issues are addressed by the system operators separately from market clearing through the sale and purchase of ancillary services or balancing actions.

In the US markets, congestion management is integrated with energy market clearing both in the day-ahead and real-time markets. The market-clearing mechanism provides a solution that meets demand at the lowest cost while satisfying all network flow constraints. In the event of congestion, the market-clearing prices differ across network locations (see further discussion on network congestion issues in Chapter 4).

2.4.3 Arbitrage between Day-ahead, Intraday and Real-Time Markets

In Europe, market participants are typically discouraged from taking speculative positions in the day-ahead or intraday markets, and close them on the real-time market. Such positions involve running voluntary imbalances in real time and paying imbalance prices. As we discussed in Section 2.3.1, imbalance prices often feature the dual-price system that makes such arbitrage unprofitable regardless of the market outcome. As a result, in the European approach arbitrage between the forward and real-time prices is prevented or considerably limited.

In the US, however, the real-time market and imbalance settlement are very tightly related. Imbalances are settled at real-time prices regardless of the direction of the imbalance. Unlike the European system, imbalance prices are locational and reflect the real-time cost of network congestion. Thus the incentives to run imbalances are properly aligned both on a system and on a locational level.

Furthermore, the US markets allow purely speculative purchases and sales in the day-ahead market that are settled at real-time prices. These speculative positions can be achieved through virtual bids submitted in the day-ahead market, that is, bids that are not associated with physical generating or consumption assets. Such arbitrage between the day-ahead and real-time markets results in day-ahead prices converging in the long term with real-time prices. As a result, day-ahead prices more accurately reflect the real-time market situation.

2.4.4 Integration of the Energy and Operating Reserve Markets

In the European model, energy is traded and reserve capacity is procured in separate market venues, running one after the other. As discussed above, this solution places the risk on the generators, since assessing the most profitable offer price for one product requires forecasting the price of the other. Greater risk for the suppliers may translate among other things into higher energy prices and reserve procurement costs. Furthermore, forecast errors may then cause inefficient offer and production decisions.

In the US model, however, the spot energy market and the operating reserve market are cleared simultaneously. The spot market-clearing algorithm minimises the cost of matching load under the constraint that enough spare generation capacity be available to provide reserve at all times. The market-clearing algorithm calculates the clearing price of energy as well as of each type of operating reserves. Each MW scheduled for production receives the energy clearing price and each MW supplying operating reserve is paid the clearing price of the operating reserve. The

clearing price of operating reserves represents the marginal opportunity cost of maintaining the required amount of spare capacity.

NOTES

1. In fact, the value of electricity can vary widely from one minute to the next. For the purpose of trading, however, the standard electricity product is commonly defined in terms of total production, or consumption, during a fixed hour.
2. Examples of power exchanges operating on the spot and derivative markets are APX-ENDEX and N2EX in the UK, and APX-ENDEX in the Netherlands and Belgium; EPEX-SPOT and EEX operate, respectively, the spot and derivative markets in Germany and France.
3. See next subsection where we discuss complex offers.
4. If a vertical segment of demand and a vertical segment of supply overlap, the market-clearing price is undetermined. In that case, selection of the clearing price follows conventional rules that may vary in different markets.
5. In this section we draw heavily on Baldick, R., 2009. 'Single Clearing Price in Electricity Markets'; Report, February, (available at: <http://www.competecoalition.com/files/Baldick%20study.pdf>), which provides a comprehensive survey of the debate on the desirable auction format for wholesale electricity markets.
6. Pay-as-bid is not implemented in spot electricity markets. However, it is implemented in some European balancing markets; see Section 2.3.
7. Fabra, N., Von der Fehr, N. and Harbord, D., 2006. 'Designing Electricity Auctions', *RAND Journal of Economics*, **37**(1), 23–46; Federico, G. and Rahman, D., 2003. 'Bidding in an Electricity Pay-as-Bid Auction', *Journal of Regulatory Economics*, **24**(2), 175–211; Bunn, D. and Bower, J., 2001. 'Experimental Analysis of the Efficiency of Uniform Price versus Discriminatory Auctions in the England and Wales Electricity Market', *Journal of Economic Dynamics and Control*, **25**(3–4), 561–92; Rassenti, S., Smith, V. and Wilson, B., 2002. 'Discriminatory Price Auctions in Electricity Markets: Low Volatility at the Expense of High Price Levels', *Journal of Regulatory Economics*, **23**(2), 109–123; Cramton, P. and Stoft, S., 2007. 'Why We Need to Stick with Uniform-Price Auctions in Electricity Markets', *Electricity Journal*, **20**(1), 26–37.
8. We are abstracting here from the impact of the generator's inter-temporal constraints on the optimal bidding strategy. We discuss how these constraints are addressed by different market types later in this chapter.
9. These measures are discussed in Chapter 5.
10. The available generation capacity can be used either to produce electricity or to provide operating reserve. Therefore, generation scarcity must be assessed by comparing the available capacity with the sum of energy and operating reserve. For the sake of simplicity of exposition, in this section we refer only to the demand for energy. In Section 2.3.3 we address the relationship between the market-clearing prices for electricity and for operating reserve.
11. Note that we assess market demand and supply with reference only to physical resources. In other terms our (competitive) supply function is the system generators' variable cost function, and our demand function represents power withdrawals from the delivery points. In most European markets, bids and offers in the spot markets are less physical, as they refer to the net portfolio position of the market participants. This explains why in those spot power markets price-dependent demand bids are commonly observed. To the extent that there is perfect foresight and no impediments to trading, this feature of market design does not impact on the market outcome (quantity injected and withdrawn and spot prices), which will ultimately reflect the fundamental physical conditions of demand and supply.

12. Some large consumers have the capacity and may find it profitable to reduce consumption in response to price spikes. Typically those consumers find it more profitable to sell their availability to reduce consumption at short notice as an ancillary service or in the balancing market (see Section 2.3). In Figure 2.5 one may think of the consumption by those consumers in normal conditions as concurring to determine the price-inelastic demand function. The price-dependent load reduction by those consumers may instead be represented as generation capacity.
13. At this stage, disconnection is still a possibility. Whether disconnections will be implemented or not depends on the actual system conditions in real time.
14. Council of European Energy Regulators – CEER, 2005. *Third Benchmarking Report on Quality of Electricity Supply*.
15. Note that the problem has the same nature as the one caused by the pay-as-bid market-clearing rule.
16. The eight-hour block would have higher average costs since the start-up cost would be spread over a smaller quantity.
17. The power exchange operating the derivative market in France and Germany (EEX) also functions as clearing for OTC forward trading (the volume traded in 2010 was around 700 TWh).
18. Electricity futures are sometimes referred to as ‘two-way contracts for differences’, and options as ‘one-way contracts for differences’.
19. As an alternative to generation or consumption, market participants may fulfil their delivery and consumption obligations by scheduling exports or imports with neighbouring countries.
20. In most markets, if a market participant fails to offset its contractual position with physical notifications, the system operator closes the market participant’s position on its behalf.
21. The supply contracts to end-consumers are not regarded as transactions at a wholesale level. The client’s consumption adds to the retailer’s (negative) physical position. The retailer matches that position either by notifying own production, as the utility in our example, or by a purchase on the wholesale market, as the retailer in our example.
22. Notice that the trader’s contract and physical positions are identical to those of a power exchange, since the same total quantity is bought and sold during each market session.
23. For example, the efficiency of thermal generator efficiency is affected by the ambient temperature.
24. If imbalances are assessed with reference to the sum of all production and consumption in the market participant’s portfolio there is no need to require notification of a separate production or consumption programme per each generating unit owned by the market participants or for each consumer supplied by the market participant. However, this information becomes relevant if imbalances are assessed at a deeper level of granularity and if imbalances in opposite directions are settled at different prices, as we discuss later in this section.
25. The system operator buys and sells electricity on the real-time market continuously, based on the condition of the system at a given time. It is possible that during an hour the system imbalance may change from negative to positive and vice versa, causing the system operator to buy additional production and then to buy reduced production, for example.
26. See Section 2.3.4 and Chapter 4.
27. Since the British real-time market operates under a pay-as-bid pricing rule, the prices of multiple transactions that take place in the same 30-minute interval need to be aggregated into one System Buy Price and System Sell Price. The System Sell Price is computed as the average of the 500 MWh highest-priced offers accepted in the time interval required for assessing the imbalances (a fixed half-hour in the British system). The System Buy Price is obtained as the average of the 500 MWh lowest-priced bids accepted in the 30-minute time interval.
28. Vandezande, L., Meus, L., Belmans, R., Saguan, M. and Glachant, J.-M., 2010. ‘Well-functioning Balancing Markets: A Prerequisite for Wind Power Integration’, *Energy Policy*, **38**(7), 3146–54.

29. More recent meters installed at residential consumers' homes record the volume of energy consumed in given time bands. For example, consumption occurred between 8.00 hours and 22.00 hours and between 22.00 hours and 8.00 hours. The roll-out of more advanced smart meters enabling real-time metering is being widely discussed.
30. For reasons of simplicity, we abstract, a number of issues that complicate the mechanisms actually implemented. These include, for example, the need to account for non-hourly metered generators, the fact that non-hourly meters are not all read simultaneously, the possibility of measuring the aggregate hourly consumption for each group of non-hourly metered consumers connected to each branch of the transmission network, and the implementation of a differentiated sharing factor, at different times of the day, in order to reflect the typical time profile of consumption by different categories of customers.
31. The benefit of the consumption shift would be split between all the suppliers of load-profiled customers, as the cost of the total load would decrease.
32. For reasons of simplicity, unless otherwise stated, our presentation assumes that only generators provide balancing services.
33. A call in the balancing market to increase or to reduce output is considered a change in the contractual position of the generator from its position at gate closure. As a consequence, a unit that does not achieve the production level after its balancing offer has been accepted incurs imbalance charges.
34. Arbitrage through voluntary imbalances can be limited by the structure of the imbalance prices in order to encourage market participants to balance their position at gate closure, as we discuss in Section 2.4.
35. Major electricity consumers may also supply balancing provided they are capable of reducing electricity withdrawals at the system operator's instruction.
36. Generators providing secondary reserves also need to be equipped with devices allowing production to vary automatically in response to a signal sent by the system operator.
37. In this case each offer on the capacity market also specifies the price of electricity if the capacity is activated. If the offer is accepted then that energy price will be offered on the balancing market. Primary reserve injections and injection reductions typically balance out over very short timeframes; such injections are typically not remunerated. The price for the electricity produced when the secondary reserve capacity is activated can either be fixed when the reserve offers are selected, or set as equal to the price prevailing on the balancing market in real time.
38. In the case of reserve capacity for downward regulation, a generator's energy schedule needs to be above minimum output, thus providing foot room.
39. The alternative approach is mentioned in the next section and analysed fully in Chapter 4.
40. Loads may also be re-dispatched, even though that is typically not the least-cost solution available to the system operator.
41. Real-world transmission networks often have a complex topology featuring multiple loops and parallel paths connecting any two points in the network. In such networks, binding network constraints may create relations among transactions across different locations with various degrees of complementarity or substitution. For example, if a transaction between two nodes is limited by a transmission constraint, this constraint can be relieved to a different degree by increasing net injections at some nodes of the network or by decreasing net injections at some other nodes. See Chapter 4.
42. The detailed approach and criteria for bid/offer exclusion from the balancing price calculation is described in the System Management Action Flagging Methodology (SMAF).