3. Generation capacity adequacy

Guido Cervigni, Andrea Commisso and Dmitri Perekhodtsev

3.1 INTRODUCTION

In the market design discussed so far, known as an 'energy-only' market, generators obtain revenues (only) from selling electricity and ancillary services. In this context, persistently high electricity and ancillary service prices are relied upon to attract investment in generation capacity when the existing capacity is below the equilibrium level. Conversely, low electricity and ancillary service prices discourage capital accumulation when installed capacity is above the equilibrium level.

As in most other markets, the level of installed production capacity in energy-only markets is determined by the interaction between demand and supply of the final products supplied. This differs substantially from the traditional approach, in which utilities meet reliability and resource adequacy requirements according to engineering standards regarding the acceptable hours of load shedding, based on the expected load variance and generator availability. In the energy-only design, market forces rather than engineering standards determine the installed capacity, and ultimately the level of reliability.

A pure energy-only market design is difficult to implement for both political and technical reasons. First, an energy-only market is characterised by generally moderate energy prices with rare price spikes. Occasional capacity shortages and extremely high scarcity-related prices are a normal feature of a well-functioning energy-only market. Policy makers and regulators are generally unwilling to accept the potentially severe price spikes and the instances of demand rationing (which may include rolling blackouts) associated with energy-only markets.

Second, wholesale electricity markets are particularly vulnerable to market power when existing capacity is close to full utilisation. Market-power mitigation mechanisms may intentionally or unintentionally reduce the price for electricity and ancillary services during conditions of scarcity. If this happens, the optimal level of generation capacity cannot operate profitably.

Finally, in energy-only markets a large portion of some generators' fixed costs is covered by the revenues obtained during genuine conditions of scarcity. This makes the investment in generation capacity risky, since even small changes in the number of scarcity events can have a dramatic impact on the producers' revenues. The problem is exacerbated by the fact that in the event of scarcity the price needs to be set administratively, because electricity demand is largely insensitive to price. To some extent even the detection of situations of scarcity may be impaired by some features of the market design.

These issues, combined with the long construction times of generation plants, result in a boom-and-bust pattern in generation investment, and governments and regulators are concerned that during the low phases of investment cycles installed generation capacity may not be sufficient to match the load at all times.

Almost all power markets implement forms of out-of-market backstop mechanisms in order to ensure reliability and sufficient generation capacity to match load. In addition, explicit capacity support schemes have been implemented in the US markets, and in Spain and Italy in Europe. Their introduction is currently under discussion in the UK, France and Germany.

In Section 3.2 we assess how the specific features of electricity impact on the economic mechanism driving investments in generation capacity, and investigate why this may create the need for capacity support schemes. In Section 3.3 we analyse alternative capacity support schemes.

Throughout the chapter we maintain the assumption of perfect competition in the electricity and ancillary service markets; we specifically assume that entry and exit on the market are frictionless. This ensures that – at equilibrium – investment in generation capacity yields no more than the minimum rate of return necessary to attract capital to the industry. Under this assumption, capacity support schemes do not increase the return of investment in power generation, but only the level of installed capacity.

3.2 THE RATIONALE FOR GENERATION CAPACITY SUPPORT SCHEMES

In this section we assess how the specific features of electricity impact on the economic mechanism driving investment in generation capacity, and investigate why this may create the need for capacity support schemes.

Capacity adequacy concerns mostly relate to distortions in the market outcome under conditions of scarcity. Scarcity hours are particularly important in the electricity industry because a large portion of some

generators' fixed costs must be recovered during these hours. In fact the generating unit with the highest variable cost installed in the system, the marginal unit, should cover its entire fixed cost by producing during scarcity hours, as it is only during these hours – under perfect competition – that the market price is greater than the variable cost of the marginal generator. For this reason, even moderate distortions of the electricity prices prevailing during scarcity hours, or in the number of scarcity hours, could have a major impact on the generators' profitability.

We identified three broad motivations supporting the introduction of capacity support schemes. First, a large part of electricity demand is currently price inflexible in the short run. When the price-insensitive portion of demand exceeds available generation capacity, involuntary load reduction via disconnections, or load shedding, may become necessary, as we discussed extensively in Section 2.2.1. In this case the price for electricity must be administratively set. Imperfections in the administrative process that sets the market price for electricity in the event of scarcity may bias the incentives to invest in generation capacity.

Second, some features of the market design and regulatory system may prevent energy and operating reserve prices from rising to levels that correctly reflect conditions of scarcity. In this case the generation capacity is under-remunerated in scarcity situations, which results in underinvestment.

Third, capacity adequacy concerns are sometimes motivated by the specific risk structure of the generation business, such that small changes in demand or supply conditions can have a dramatic impact on generators' profitability at times of scarcity. While the first two issues call for mechanisms that integrate the generators' income in order to attract an efficient level of investment, the third issue can be handled by coordinating the timing of investments in generation capacity in order to reduce the risk for investors. A more certain environment is expected to reduce the rate of return required by investors, to the ultimate benefit of consumers.

We discuss possible flaws of the price-setting mechanism in the event of scarcity in Section 3.2.1, the 'missing money' problem in Section 3.2.2 and the coordination role of capacity support schemes in Section 3.2.3.

3.2.1 Flaws in the Assessment of the Value of Lost Load

Figure 3.1 shows the wholesale electricity market equilibrium for two sets of hourly demands with different price elasticities. The price-insensitive demand results in fewer scarcity hours, that is, hours when the market–clearing price rises above the system's marginal cost (SMC), the marginal cost of the most expensive generating unit in order to ration demand.

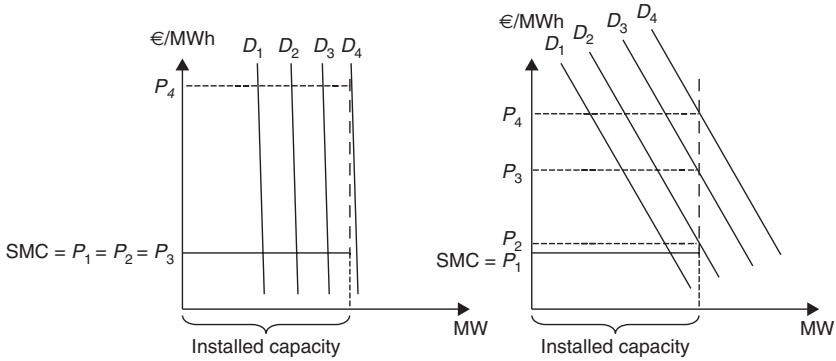

Figure 3.1 Market equilibrium with different demand elasticity

However, during those hours the distance of the clearing price from the system's marginal cost is greater than it would be if demand were more flexible.

In the current situation a large portion of the electricity demand is completely price insensitive in the day-ahead to real-time timeframe. This reflects the technical features of most of the metering systems currently in place. The meters installed at small consumers' premises typically record the consumers' total withdrawal over long periods, typically one month. Older meters record only the total consumption since the equipment was installed. It is therefore impossible to measure the volume of electricity withdrawn by a consumer during each hour. When hourly consumption is not known, retail prices cannot reflect wholesale market prices, and therefore cannot signal to consumers the cost of their consumption in each hour.

Without demand response, no matter how high (wholesale) prices rise, load will not reduce to the level of available generation capacity during scarcity events. Therefore a regulatory solution for scarcity pricing and involuntary load reductions must be implemented in order to avoid uncontrolled widespread service disruptions. Given that selective disconnection is technically infeasible, if necessary all the consumers connected to the same network branch will be disconnected at the same time.[1]

As such quantity-rationing events have no associated market-clearing price, the price for electricity in scarcity hours must be set administratively. The appropriate regulated price in such circumstances is the estimated value of lost load, or VoLL. The VoLL is based on an estimate of the amount that customers would be willing to pay in order to avoid being disconnected. In other words, the VoLL is supposed to be the price that makes consumers indifferent between consuming electricity at that price

and not consuming. The VoLL is typically found to be several orders of magnitude greater than average electricity prices. VoLLs in the range of €5,000–€10,000 are commonly regarded as plausible.

Implementing VoLL pricing and load curtailment is not without problems. First, load curtailment is perceived by end–consumers as being unfair. Curtailed consumers typically do not receive payments equivalent to the VoLL from their suppliers, while non-curtailed consumers are not charged for the VoLL.

Second, although each consumer might give a different value to electricity, current technology makes it impossible to selectively disconnect consumers based on their individual valuation of electricity. It is then impossible to provide incentives to consumers to reveal their individual valuation.

Finally, load shedding and price spikes rapidly become a matter for political concern.

For these reasons governments and regulators and system operators tend to pursue more or less explicit capacity targets, rather than relying on extremely high prices under conditions of scarcity in order to attract investment in generation.

3.2.2 The Missing Money Problem

The missing money problem occurs when some elements of the market design, industry regulation or industry practices cause generators' revenues to be systematically insufficient to attract the efficient level of investment. When revenue deficiency becomes a structural feature of the market, the result is a drop in installed capacity.

Capacity support mechanisms are therefore intended to integrate generators' income. Below we discuss the potential causes of missing money.

Market-power mitigation measures

As we discuss in detail in Chapter 5, wholesale electricity markets are particularly vulnerable to the exercise of market power when existing capacity is close to full utilisation. When demand approaches the level of available capacity, even relatively small generators enjoy market power. Since both electricity supply and demand are to a large extent price inflexible, withdrawing even a small amount of capacity from the market when the system is tight can be very profitable for a generator, as it may result in a dramatic increase of the market-clearing price. This happens especially if capacity withdrawal results in a scarcity situation, that is, if the market-clearing price jumps from the system marginal cost to the much higher VoLL.

A straightforward solution to reduce generators' incentive to withdraw

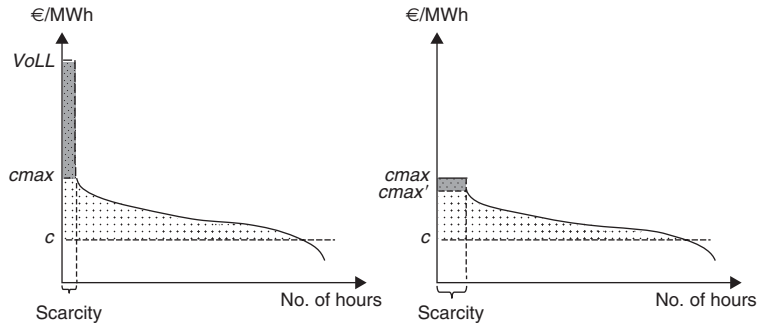

Figure 3.2 Price duration curves

capacity when the system is tight consists of setting the scarcity price at a level much below the VoLL. For example, the scarcity price could be administratively set as equivalent to the marginal cost of the most expensive generating unit activated. With such a price cap, the expected revenue for the generators is below what is necessary to attract an efficient level of investment. This situation is illustrated in Figure 3.2, where we show a price duration curve. The price duration curve shows, for each price level, the number of hours the market-clearing price is above that level.

In the figure, the area with the grey background represents the profit that 1 MW of generation capacity with variable cost *c* obtains on the wholesale market. The profit is equivalent to the sum of the difference between the market price and the generator's variable cost in each hour. Over the lifetime of the generator, capacity settles to the efficient level and composition, and market prices are such that the efficient capacity level obtains the standard return on investment.

If, during conditions of scarcity, the price is set to be equivalent to the most expensive generator's cost (*cmax* in the figure) instead of the VoLL, the generators' profits are reduced by the darker area.[2] Note that because of the price cap, the most expensive unit does not receive any contribution to fixed costs from selling energy. In the long run, standard profitability conditions are re-established via the entry/exit process. Capacity settles at a lower level as the units with variable cost *cmax* are not replaced, and the new system marginal cost becomes *cmax*'. This means that the number of scarcity hours increases until the new dark area is large enough to cover the generators' fixed costs,[3] as shown in the right panel of Figure 3.2.

When the scarcity price is below the VoLL, a mechanism integrating the generators' revenues is necessary to ensure that the efficient capacity level is available in the system at all times.

In most markets, price caps below common estimates of VoLL are

imposed. This happens, for example, in Nordpool (the Nordic power market), in the Australian day-ahead market and in the US in ERCOT (the power market operating in Texas). The Australian market and ERCOT also have additional price mitigation measures which limit the duration of elevated scarcity prices. If prices remain above a predefined threshold for a certain period of time a price cap is enforced in Australia, and the normal price cap is lowered in ERCOT.

An alternative approach to market power mitigation, discussed in Chapter 5, Section 5.3, involves capping the generators' offers only at times when they are considered to enjoy significant market power. However, when the system is tight it is very hard to distinguish between high prices that reflect a genuine situation of scarcity and high prices that are the result of exercise of market power. This was particularly evident in the aftermath of the 2000–01 California power crises where, according to some observers, high loads and low water availability for electricity production, combined with market manipulation, resulted in extremely high prices. Market-power mitigation mechanisms based on selective capping of the generators' offers may also be activated in situations of genuine scarcity and create revenue deficiencies for the generators.

For this reason, such measures are typically associated with capacity support mechanisms providing an additional source of revenues to the generators. The class of capacity support schemes presented in Section 3.3.1 accomplishes both functions: limiting generators' revenues during scarcity events, and making up the missing money with a payment for capacity availability over a longer period.

Out-of-market procurement of reserve services

Almost all power markets have developed out-of-market backstop mechanisms for ensuring reliability and sufficient capacity. In most cases, market operators simply procure reserve capacity outside the market framework if they expect peak capacity to be short of their targeted reliability standard. In Nordpool, for example, when capacity is forecast to be insufficient on a three-year forward basis to meet the reliability target, the system operator is authorised to procure peaking resources under long-term contracts with the costs of the procurement paid for by the state.[4] In the UK, operating reserve is procured by the system operator under long-term contracts of up to five years in order to provide sufficient investment signals to providers and allow enough time for the repayment of a provider's investment.[5] In some US markets, out-of-market capacity purchases have been made in the form of reliability-must-run contracts, which are intended to retain in the system capacity resources that might otherwise be retired or mothballed.

Such backstop measures may displace market-driven investment in generation capacity and inhibit the development of demand-response measures in energy-only markets if they prevent market prices rising to VoLL during scarcity events. As we discuss in Section 3.3.3, the distortions caused by out-of-market mechanisms can be overcome by appropriately designed scarcity pricing rules, administratively setting the market price to the VoLL each time the backstop resources are activated.

Lack of transparency

A number of markets do not implement scarcity pricing rules that administratively set the electricity price to VoLL when scarcity conditions are detected by the system operator. These markets rely on generators to increase their bid prices above marginal costs in order to set scarcity prices. For this mechanism to be effective, it is crucial that generators are in a position to correctly anticipate scarcity situations. Lack of information about the demand and supply may cause some scarcity situations to go undetected by the market participants, exacerbating the missing money problem.

When multiple related markets are cleared independently, the profit-maximising bid for a generator in one market depends on the expected equilibrium price in the others. Arbitrage between the related markets should result in all the markets clearing at a price consistent with the overall demand and supply. As discussed in Chapter 2, some elements of the designs implemented in Europe may make arbitrage between the various markets difficult: energy and reserve markets are cleared independently; the design of the day-ahead, intraday and real-time markets is not always homogeneous; and speculative trading against real-time prices is limited in some markets. Imperfect arbitrage may result in scarcity conditions not being reflected in the same way in different markets. Consider, for example, the case in which the energy market clears first and the operating reserve capacity market clears later. If the generators fail to anticipate the scarcity situation when formulating their bids in the energy market, the electricity clearing price will not signal scarcity because the available generation capacity is greater than the demand in the energy market. However, scarcity conditions will emerge – and result in high prices – in the operating reserve market. If this is the case, the generation capacity committed on the electricity market will receive a price that does not correctly reflect its value.

3.2.3 Coordination of Investment Decisions

The high level of risk in generation investments is sometimes mentioned as a reason for the introduction of capacity support mechanisms. In this

respect, the relevant feature of some capacity support schemes is that they coordinate market participants' investment decisions. To the extent that such coordination reduces the uncertainty faced by generators, it also reduces the required rate of return on the investment in generation capacity. As a consequence, all other things being equal, a higher level of capacity will be installed.

In order to illustrate this approach, consider an investor assessing the opportunity to invest in a 1,000 MW plant to be in service in year *t*. The profitability of this investment crucially depends on the decisions of other potential investors to enter the market. If just two new plants were brought into service in year *t* instead of only one, electricity prices might turn out much lower than if only one were built. The impact of the second project on market prices could be considerable for several years, potentially undermining the profitability of both projects.

Although an efficient investment pattern would include 1,000 MW additional capacity in service from year *t*, each investor will want to reduce its investment's vulnerability to other investors' decisions. Such a strategy can be expected to lead to investments being delayed when compared with the efficient path. If this happened, an inefficiently low level of available capacity would be available in year *t*.

In this context, a mechanism that coordinates investment decisions could be beneficial. A central entity could set a capacity target for each year, and select the parties that would make the target capacity available in exchange for payment. Selection of the new capacity provider could take place by means of auctions.

In our example, the target level at time *t* would be such that only the additional 1,000 MW would receive the capacity payment. If the central entity commits to paying for capacity availability well in advance, the scheme coordinates the decisions of the potential investors. Only one of the potential investors obtains the capacity payment for 1,000 MW capacity at time *t*. The other investors would be unlikely to sink money into making additional capacity available at time *t*, as they know that such an investment would lead to excess capacity overall, and would therefore be unprofitable.

Here the capacity support scheme is welfare improving because it coordinates the timing of the investors' decisions. While in the previous section the capacity support scheme increases the generators' income in order to compensate for the effects of the market flaws, here the capacity support scheme acts mainly as a coordination device. In our example, very little or no compensation would be required by the winner of the auction, if the central entity auctions off only 1,000 MW of incremental capacity. In this case, each of the potential investors would find it profitable to make

1,000 MW capacity at time *t* with no further compensation, provided that no more than 1,000 MW new capacity is built in total. What the auction process delivers is mainly the certainty as to who will make the investment. If instead, as typically happens, the capacity target pursued by the central entity is greater than the level that would attract the investment, then the auction will clear at a higher price.

Note also that the reallocation of risk from the generators to the central entity acting on behalf of the consumers is a byproduct of this measure, not the source of its expected welfare-improving effect.

3.3 CAPACITY SUPPORT SCHEMES

Three broad approaches to the design of capacity support schemes can be identified. The first approach sets the price that a central entity, on behalf of the consumers, commits to pay for all available capacity. Capacity payments add to the revenues obtained by generators from selling electricity and ancillary services; the higher expected income is supposed to attract additional investment in generation capacity. The second approach sets the volume of available capacity that the central entity commits to paying for, either directly or by placing an obligation on the load-serving entities. This creates the demand for a product, the available capacity, which generators can supply. The interaction between the regulatory-driven demand and the supply of available capacity determines the market-clearing price for the available capacity. The third approach consists of reserving a certain generation capacity for use only in scarcity situations, as a substitute for load curtailment.

In the rest of this section we discuss each approach in turn.

3.3.1 Capacity Payments

Capacity payments are administratively set payments per MW for available capacity, regardless of whether it is dispatched to run. Capacity payments are intended to provide generators with additional revenues equivalent to the missing money. Consider, for example, a market where the missing money issue is created by an overall price cap equivalent to the variable cost of the marginal generator, typically an open-cycle gas-fired unit. As illustrated in Section 3.2.2, the efficient level of capacity would be under-remunerated because of the cap. Each MW of installed capacity would, over its lifetime, miss out on revenues equivalent to the fixed cost of the marginal unit, the variable cost of which sets the price cap. In this case the efficient capacity payment would equal the fixed cost of the open-

cycle gas-fired unit. This would neutralise the impact of the price cap on the generators' income,[6] and therefore re-establish the correct investment incentives. The level and allocation of capacity payments is a matter of administrative judgement.

Different schemes feature different availability requirements. For illustration purposes we shall consider two extreme methods. The first methodology would pay 1/8,760 of the annual fixed cost of the marginal unit for all capacity that turns out to be available during each hour of the year.[7] The drawback of this approach is that it does not provide incentives to make capacity available when the system needs it most; the generator has the same incentive to be in service all hours, independently of the level of demand.

The second extreme methodology would pay $1/N$ of the annual fixed cost of the marginal unit for all the capacity that turns out to be available in each the N hours of the year when the system operator expects scarcity. The advantage of this approach is that it provides stronger incentives for generators to make capacity available when the system needs it most. If the N hours selected by the system operator, and only those N hours, turn out to be scarcity hours, this methodology provides exactly the same incentives that would be provided by pricing electricity at the demand-rationing VoLL price. The scheme, however, has an additional advantage over VoLL scarcity pricing: it removes the burden of predicting when scarcity will occur from generators. For example, they can plan maintenance outages at times other than those when the capacity payment is granted. The drawback of this approach is that it depends on the system operator's ability to predict exactly when scarcity conditions will occur. Typically, critical system conditions could manifest 5–20 hours per year, and predicting, say, a year in advance when those hours will be is a difficult exercise.

The trade-off between the power of the incentives to make capacity available and the risk of excluding some scarcity hours from the scope of the mechanism is addressed in practice by granting capacity payments in exchange for availability for a relatively large subset of a year's hours, for example one or two thousand, when demand is expected to be high. In Chile, for example, capacity payments are granted for availability in the months May–September; in Colombia, in the dry December–April season when hydropower production is limited.[8] In Italy the set of critical days when capacity availability will be remunerated is set yearly by the system operator.

Administratively determined capacity payments can target new resources. The Spanish capacity payment system, for example, comprises a component granted only to new capacity. This approach is sometimes supported in policy discussions, because it reduces the initial amount

of capacity payments compared with capacity payments being granted for the entire capacity. However, it is distortive, since the new capacity attracted by the selective capacity payment will exacerbate the missing money problem for existing generators that do not receive the capacity payment. This will accelerate substitution of the generating fleet.

The British pool system, operational between 1991 and 2000, featured a capacity payment component that was paid to all generators available for dispatch, irrespective of their activation. The capacity payment was computed the day before delivery as the expected value, over the probability distribution function of the demand realisations, of the scarcity rent on the day of delivery. As a result, capacity payments would be low when available capacity was high compared with load, and payments increased as the reserve margins shrank.[9] The design of the early UK capacity payment mechanism can be interpreted, more than as a capacity support system, as a way to implement a pure energy-only market under the constraint that the real-time price for electricity and operating reserve be fixed one day in advance. However, the parameters of the mechanisms were set in a way such that it resulted in generous payments to the generators.[10]

3.3.2 Capacity Requirements

With capacity payments, available capacity levels remain uncertain, since they depend on the market response to administratively set prices. An alternative approach is to ensure resource adequacy by imposing a reserve margin requirement on all electricity retailers. In this section we discuss two support mechanisms based on capacity requirements that differ in the content of the obligation placed on capacity suppliers.

Reserve requirements

In this approach the system operator sets the capacity requirement. The required level of installed capacity is generally set around 115–118 per cent of the peak load, a figure derived from engineering standards. Like the level of the capacity payments, reserve requirements are set administratively, and the trade-off between the cost of achieving the reliability target and the value provided by that reliability is typically not explicitly addressed.

The reserve requirement is then divided between retail suppliers in proportion to the expected contribution of their clients to peak load. Each retail supplier is responsible for acquiring capacity rights that exceed its predicted peak load by the required reserve margin, either through self-supply or by contracting available capacity from generators. The capacity availability contracts can be negotiated either bilaterally or on organised markets.

By selling capacity, a generator commits to make the contractual volume of capacity available during the period of the contract. The obligation is fulfilled independently of the actual use of that capacity. The generator can use that capacity to deliver electricity sold bilaterally, on the spot market or on the real-time market. Even if the capacity turns out to not be used, the obligation has been fulfilled so long as it has been offered as operating reserve and on the real-time market. Financial penalties apply if the capacity is not delivered. Resources wishing to supply capacity apply to the system operator to be assigned capacity credits, typically based on their historic availability record; underdelivery can also be penalized through a reduction in the capacity credits assigned in the future.

Placing a capacity requirement on retailers creates the demand for capacity that meets the generators' supply. A market for capacity is then established. The price in that market settles at the level that attracts investment in generation capacity up to the system operator's requirement. Under the usual perfect competition assumptions, generators' revenues from selling capacity availability are equivalent to the missing money.

Mechanisms based on capacity requirements are extensively implemented in US markets. Earlier capacity requirement systems were implemented in the context of traditionally regulated markets, where integrated utilities carried a regulatory obligation to procure the generating capacity needed to meet the resource requirements in their (exclusive) service areas. The absence of retail competition allowed the utilities to recover the costs associated with that obligation through regulated retail rates. That meant that the need for adjusting the utilities' capacity portfolios via trading was limited to transitory imbalances, since the resource planning requirements were overseen or enforced by the state regulators.

In restructured markets, with retail competition and small retailers, some features of the older mechanisms have caused concern. In particular, enforcing the capacity requirement just days or months before the relevant delivery period may lead to extreme price volatility, with capacity prices jumping from the cap (when there is insufficient capacity) to zero (when there is excess capacity). This happens because in such a short time horizon both the demand for and the supply of capacity are highly inelastic. The supply of capacity to deliver within a period of months does not include units not yet built. As a consequence, the entrant's cost does not act as a ceiling to the prices of capacity.

In addition, suppliers and possibly buyers of capacity contracts may enjoy significant market power when the system is close to the target resource requirement: suppliers may be able to move price from close to zero to the cap even by withholding relatively small amounts of capacity. Buyers may similarly be able to move capacity price levels close to zero by

slightly reducing their demand for capacity, for example by declaring that they will meet part of their obligation via self-supply. Finally, if capacity deficiencies are detected only slightly in advance, it may be impossible or extremely costly to the system operator to make up the missing resources.

In order to address those concerns, forward reserve requirements have been introduced in several US power markets. For example in PJM (the power market in Pennsylvania, New Jersey and Maryland) and in ISO-NE (the power market in New England) load-serving entities are required to procure sufficient resources for up to three years ahead of the delivery year. In addition, both PJM and ISO-NE allow some suppliers of new capacity to lock in capacity prices for three to five years.

Requiring resource commitments sufficiently in advance of delivery leaves enough time for either market participants or the system operator to procure additional resources if deficiencies are detected. The time horizon also gives capacity suppliers enough time to modify their resource development plans, for example by bringing mothballed plants back online, making the capital investments necessary to defer the retirement of other plants, speeding up the development of a new power plant, or developing additional demand response capabilities. As a result, the price elasticity of the capacity supply curve rises, price volatility is reduced and competition increases.

In the PJM and NYISO (the power market in the state of New York) markets a certain degree of price elasticity is also introduced on the demand side by implementing a downward-sloping capacity demand curve, which varies the resource adequacy requirements as a function of capacity prices. As we have already shown in Figure 3.1, a flatter capacity demand curve reduces price volatility, as a shift in the demand or supply curve leads to a smaller change in the market-clearing price. However, the capacity demand curve implemented in PJM and NYISO is not intended to represent the consumers' (estimated) willingness to pay for capacity. Instead, consumers are assumed to be available to pay a price equivalent to the estimated building cost of new peaking resources for an administratively set target level of capacity. Then the slope of the curve near the target level of capacity is based on an administrative judgement As a result, rather than reflecting consumer preferences, the capacity demand curve basically implements a cost-based cap on the price of capacity.

In order to address transmission congestion issues, the capacity requirements are imposed on a zonal or locational basis in some markets, including PJM and NYISO.

With retail liberalisation, the number of retailers trading capacity contracts has increased. In addition, the possibility for consumers to switch supplier creates an additional need for trading capacity contracts, in order

to adjust each retailer's capacity portfolio to the changing customer base. In order to address this development, centralised capacity exchanges run by the system operators have been introduced in several US markets. Centralised markets provide a backstop procurement mechanism, reduce transactions costs and provide greater liquidity and pricing transparency.

Energy options backed by generation capacity

Administratively set capacity targets may also be achieved through financial contracts with the suppliers of capacity. Capacity support schemes based on financial obligations are implemented in Colombia – the 'firm energy obligation'. In Europe a mechanism along the same lines is being introduced in Italy.

In this methodology, the capacity contract takes the form of a call option on the generators' capacity. In exchange for a fixed fee, the supplier of generation capacity commits to pay the counterparty the following amount in each hour of the contract period:

$$\mathrm{Max}(0, p_t^{Spot} - p^{Strike}),$$

where p_t^{Spot} is the spot price of electricity in the hour, and p^{Strike} is the option's strike price, set as equal to the variable cost of the marginal generation unit in the system. In scarcity hours, when the price rises above the variable cost of the marginal unit, the contracted generator disburses the scarcity rent for each MW of hedged capacity. The scarcity rent is equivalent to the difference between the market price and the system marginal cost, that is, the variable cost of the marginal generator.

This provides the correct incentive for generators to make the hedged level of capacity available when the capacity is most valuable to the system. Consider the situation of a generator that has sold a 1 MW capacity contract and has not made the corresponding capacity available in a scarcity hour, when the electricity price equals VoLL. The generator suffers a loss on the capacity contract equal to $p_t^{Spot} - p^{Strike}$, which amounts to the scarcity rent in scarcity conditions. The generator can offset this loss by making available 1 MW capacity in that hour; by doing so the generator sells 1 MWh and obtains profit $p_t^{Spot} - VC$, where VC is the generator's variable cost. The net profit to the generator is then equivalent to the payment due under the capacity contract:

$$-(p_t^{Spot} - p^{Strike}) + (p_t^{Spot} - VC) = (p^{Strike} - VC).$$

A further advantage of this methodology is that it mitigates market power by capping the net revenues for contracted capacity at the system

marginal cost. However, while it caps generator revenues, the mechanism does not cap the market-clearing price, which will rise to the level necessary to ration demand when the system is tight. As a result, the incentives to develop demand response resources are not distorted as they would be in the case of a price cap.[11]

In the Colombian implementation, additional provisions ensure that the financial obligation placed on the supplier of capacity is backed by physical generation capacity, and that the capacity is operated in such a way as to ensure its availability at scarcity times. For example, contracted thermal generators must provide proof of fuel availability during the commitment period.

Like installed capacity requirements, energy options backed by generation capacity may be procured directly by the system operator, or by placing an obligation on the load-serving entities. The options may have different duration, and may be procured more or less in advance of the commitment period. Finally, the option's strike prices may be fixed or indexed.

In Colombia, the system operator sets the capacity requirement and procures the corresponding energy options three years ahead of the start of the commitment period, which ranges from one to 20 years. The energy options are procured via auctions where generators are selected according to their bids for the option's fixed fee. The option's strike prices are indexed to the price of natural gas, the fuel firing the marginal open-cycle gas turbine (OCGT) capacity.

Capacity support mechanisms based on capacity requirements may contribute to coordinating generation capacity investment decisions, provided that the capacity contracts are awarded well in advance of the time of delivery. In this case a would-be investor may make the decision to build new capacity conditional on the auction's outcome. The auction outcome conveys important information to the participants. Broadly speaking, the auction's winners:

- learn that they are probably more efficient (or less risk-averse) than the others; and
- hedge part of their revenues.

Would-be investors not awarded capacity contracts in the auction learn that:

- they are probably less efficient than the winners;
- the system operator's capacity requirement will be covered by the winners of the auction; and

- in the event that they decided to go ahead with the investment, they would have no protection against excess-capacity situations, leading to too few scarcity hours.

The information conveyed by the capacity support scheme coordinates the investors' decisions. The decision to invest or not to invest is made easier. The auction's winners face strong incentives to invest, since they can be confident that the auction's losers will not invest and bring about excess capacity; the auction losers have strong incentives not to invest, since they can be confident that they will not miss a profit opportunity.

3.3.3 Reserve of Last Resort

This approach, implemented in Sweden and Finland,[12] is based on reserving part of the installed generation capacity for use only in scarcity situations, that is, as the reserve of last resort.

In order for the measure to bring about a permanent increase of installed generation capacity, the last-resort reserve has to be effectively removed from the market. This means that each time the last-resort reserve is activated, the market price for electricity (and operating reserve) rises to the VoLL, as in the event of scarcity. Otherwise, as we discussed in Section 3.2.2, the reserve of last resort will displace new capacity and the total installed capacity will not increase.

Consider, for example, the market shown in Figure 3.3, where we assume that demand and installed capacity are steady. Assume that the regulator is not satisfied that the current level of installed capacity is adequate and believes that an additional capacity of 2,000 MW is necessary.

In order to induce investment in an additional capacity of 2,000 MW, the regulator therefore contracts 2,000 MW of the existing capacity as last-resort reserve. The contracted capacity is then offered on the energy

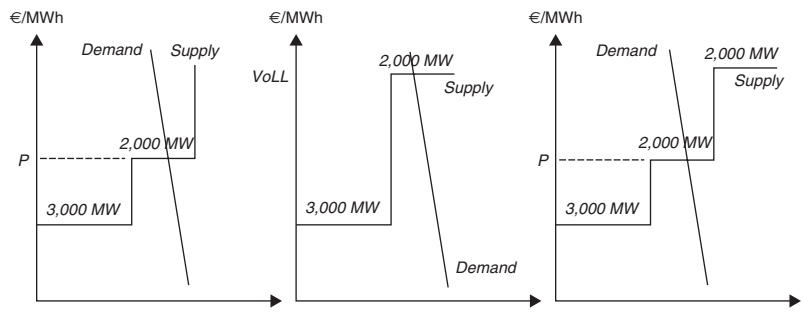

Figure 3.3 Reserve-of-last-resort scheme

and ancillary service markets at a price equal to the VoLL. The result of this measure is the market supply function represented at the centre of Figure 3.3. Consequently, in the event that the last-resort capacity is scheduled for production or to provide operating reserve the market-clearing price is the same as in the event of scarcity.

The new aggregate supply function is such that the market-clearing prices reflect scarcity conditions more often than they would without the intervention. The profitability of the existing generation capacity then increases. This attracts investment until the total installed capacity reaches the pre-intervention level, that is, until the capacity shifted to the last-resort reserve has been replaced. The new equilibrium is shown in the right panel of Figure 3.3.

The scheme based on the reserve of last resort may cause inefficiency if the units providing reserve of last resort turn out to be more efficient than some other units. In this case, cheaper units are withdrawn from the market while more expensive generators are activated to meet load. For this reason, the reserve of last resort appears particularly attractive when the regulator has the opportunity for preventing old and inefficient units being scrapped. The cost of keeping alive units that would otherwise be dismantled could be relatively low, and there would be little risk of technical inefficiency.

NOTES

1. The deployment of smart meters could in the future overcome some of the issues discussed in this section by allowing recording of hourly consumption, and therefore implementation of hourly differentiated retail prices. Alternatively, smart meters fitted with remotely controllable switches could make it possible to selectively disconnect consumers that state less willingness to pay, while charging the others the rationing price.
2. In the example we ignore the revenues that generators obtain by providing ancillary services, and we refer to electricity spot-market sessions only. These simplifications are irrelevant provided that the same price cap is consistently enforced on all services and market sessions.
3. We have assumed that the price cap has not been adjusted to the new (and lower) system marginal cost. If this happened, the installed capacity would continue to shrink.
4. See Nordel, 2007. *Guidelines for Implementation of Transitional Peak Load Arrangements: Proposal of Nordel*, available at: http://www.svk.se/Global/01_Om_oss/Pdf/Elmarknadsradet/071115NordelGuidelines.pdf.
5. See National Grid Electricity Transmission, 2008. *Long-Term Reliability Assessment, 2008–2017*, available at: http://www.nerc.com/page.php?cid=4|61.
6. The market-power mitigation effects of the cap still hold, as the capacity payment is independent of the generators' offers.
7. We ignore maintenance stops for reasons of simplicity. The correct assessment would allow each generator to obtain the annual fixed cost of the marginal unit in a number of hours equal to the difference between 8,760 – the number of hours in a year – and the duration of a standard maintenance period.

8. In Chile a penalty for failure to deliver capacity based on the VoLL re-establishes the correct incentives for the generator to make capacity available in the scarcity hours.
9. The capacity payment was computed as the product of the loss of load probability assessed by the system operator for the following day and the difference between the VoLL and the system marginal cost.
10. For a detailed analysis of this mechanism see, for example, Roques, F.A., Newbery, D.M., and Nuttall, W.J., 2005. 'Investment Incentives and Electricity Market Design: The British Experience', *Review of Network Economics*, **4**(2), 93–128.
11. A similar mechanism operates in ISO-NE, in the context of a capacity requirement system. The scarcity rent collected by the generator when the scarcity pricing mechanism is triggered is subtracted from the monthly capacity payments.
12. The RMR contracts implemented in some US markets, discussed in Section 3.2.2, are based on the same logic.