2. Optimal Wholesale Pricing and Investment in Generation: The Basics[1]

By Paul Joskow and Thomas-Olivier Léautier

1 Introduction

This Chapter presents the basic microeconomic theory underlying the formation and the structure of efficient wholesale power prices and optimal investment in dispatchable generating capacity.[2] The presentation in the Chapter is designed to be accessible to non-economists interested in understanding the basic economics of electricity supply and demand. We use examples and graphics rather than mathematics to articulate the relevant microeconomic principles. The Chapter also provides a theoretical link between the "old world" of vertically integrated regulated electricity monopolies and the "new world" based on vertical and horizontal restructuring to support competitive wholesale markets.

Over the last two decades, many countries have moved to restructure their electric power sectors to replace investment, operation and pricing of electric generation services through internal often non-transparent regulated monopoly "hierarchies" with transparent unregulated competitive wholesale market mechanisms (Joskow 1996). The conceptual basis for the design of organised wholesale electricity markets during the late 1990s and early 2000s can be traced directly to the mid-twentieth century economic-engineering literature on optimal dispatch of and optimal investment in dispatchable generating facilities and the associated development of marginal cost pricing principles for generation services. While these models were developed to apply to pre-restructuring vertically integrated electric utility monopolies subject to some kind of regulation, including government ownership, these models of generation dispatch, marginal cost pricing and investment have also guided the design of decentralised wholesale markets. That is, the basic microeconomic principles developed to facilitate efficient decisions regarding investment, generator dispatch and optimal pricing of generation services have not changed. Rather, they must now be applied to the design of wholesale markets rather than serving as guides to electric utility management and regulators governing the behaviour of vertically integrated electric power monopolies.

One of the key insights from the microeconomics of electricity production is that the structure of wholesale power prices is similar to that of other non-storable goods for which demand varies significantly across time, for example, hotels rooms or plane tickets: the price is set close to the variable cost of production when capacity exceeds demand, while it is set by the value for the marginal consumer when demand is exactly equal to capacity. For example, the price for a room at the beach on Cape Cod is close to the cost of cleanup in the winter and goes much higher in the summer. This particular price structure is called "peak-load pricing" in the power industry. The main difference between electric power and other non-storable goods is the magnitude of the peak price: the summer price may be three to four times the winter price for a room at the beach, while the peak price for power may exceed 50 or even 100 times the off-peak price.[3] Thus, while electricity supply and demand have a number of unique attributes, we can find analogies in markets for many other goods and services.

This Chapter begins with a very simple setup in Section 2 to illustrate the peak load pricing results. The model developed in this section has price responsive demand and one generating

technology. Despite its simplicity, the model yields important insights into optimal short run and long run pricing, optimal generator dispatch and optimal investment in long run equilibrium. Section 3 then introduces a number of more realistic features that also play an important role in the design of wholesale markets. These include the introduction of non-price responsive demand, an important consideration if consumers are not faced with wholesale spot prices due to metering or political constraints, demand uncertainty, customer curtailments and the value of lost load (VoLL), multiple generating technologies, transmission congestion and security of supply considerations. Section 4 concludes.

2 The Simplest Peak-Load Pricing Story

2.1 Setup

The simplest situation is characterised by a fully price-responsive electricity demand and a single production technology. These two assumptions make peak-load pricing results easy to derive and to understand. As discussed in Section 3, they are not essential: the economic intuition is unchanged when they are relaxed.

2.1.1 Demand

Units

First, a word on units. The main unit of measure for electric energy used in this text is the megawatt-hour (MWh). Kilowatt-hours (kWh) and terawatt-hours (TWh) are sometimes used. A megawatt-hour is 1,000 kWh, a terawatt-hour a million MWh. To provide orders of magnitude, average annual consumption for residential customers worldwide is about 5 MWh per year. In aggregate France consumes about 500 TWh per year and the US, a much larger country, consumes about 3,700 TWh per year. Wholesale electricity prices are usually expressed in €/MWh,[4] while retail prices are expressed in euro cents/kWh. One cent/kWh is equal to 10 €/MWh.

The rate at which energy is produced or consumed is called power. Throughout this text, we consider the hourly rate; hence it is measured in megawatts (MW), that is, megawatt-hours per hour. This is the appropriate unit for wholesale market transactions. Kilowatts (kW) and gigawatts (GW) are sometimes used. A gigawatt is 1,000 MW, while 1000 kilowatts is a megawatt. For example, peak demand for a country like France is about 92,400 MW or 92.4 GW. The peak demand in California is about 50,000 MW or 50 GW, while the (noncoincident)[5] peak demand for the entire US is about 800 GW. Kilowatts are used to measure the average residential customer's maximum demand, which is typically lower than 10 kW in OECD countries.

Load Duration Curve

Electricity demand varies greatly across hours within a year and across years. Electricity demand is higher during the day than at night and higher on weekdays than on weekends. In Northern Europe and Canada, electric heating leads to higher demand in the winter than in the summer. In most of the United States, air conditioning leads to higher demand in the summer than in the winter.

Power engineers represent this variation using a "load duration curve", which displays demand for every hour or half-hour, ordered from the highest to the lowest. As an example, the load duration curve for France in 2009 is presented in Figure 1.[6] The peak demand was reached for only one half-hour on 8 February at 7:00 pm and it was equal to 92,400 MW. The minimum demand that year was 31,526 MW.

Load duration curves are used to determine the number of hours or half-hours in which demand exceeds a given level. By construction, demand exceeds the minimum of 31,526 MW for all half-hours of the year. The curve shows that demand exceeds 60,000 MW for 5,676 half-hours of the year. These are not consecutive half-hours, some might be February evening, others might be January mid-day. This non-chronological representation of electricity demand is extremely powerful, as will be shown throughout this chapter.

Demand exceeds 90,000 MW for a very small number of the hours in the year. This means that due to the large variations in demand from hour to hour and a very limited ability to store electricity economically, a large fraction of the generating capacity installed to meet demand operates for a very small fraction of the year.

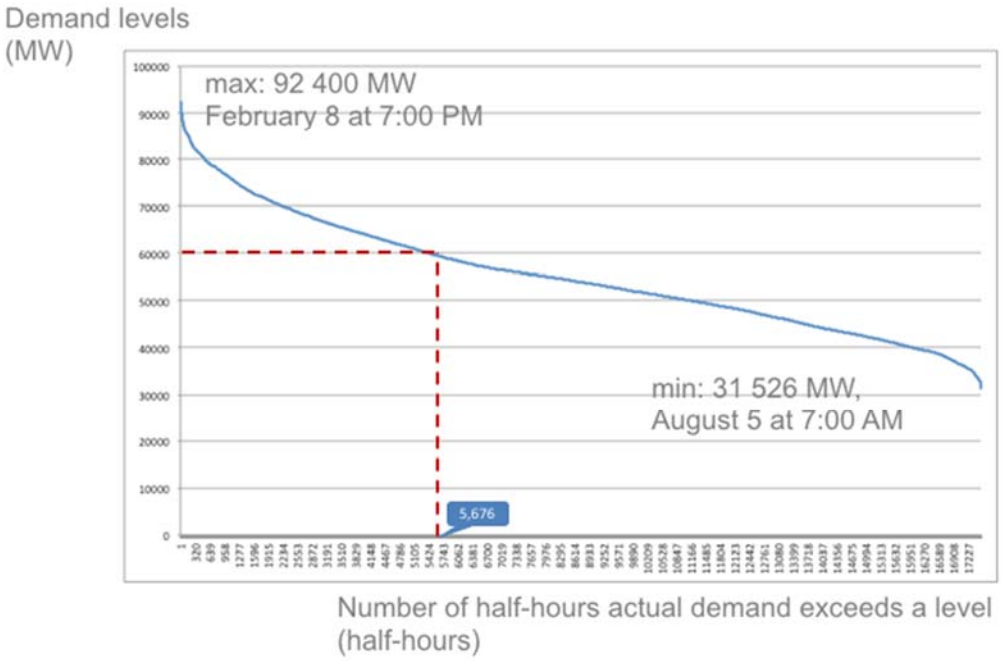

Figure 1: Load duration curve for France in 2009 Source: Léautier, 2019, figure 2.1, pg 18. All half-hours of a given year are represented on the horizontal axis; demand is represented on the vertical axis. The load duration curve represents demand for every half-hour, ordered from the highest to the lowest.

Inverse Demand Curves

In this introductory section, we assume that all consumers can adjust their consumption to respond to variations in spot prices (that is, variations in the real-time price) of electric power. When the spot price increases, consumers reduce their demand. The natural representation of demand would be to have the price on the horizontal axis and the quantity demanded at that price on the vertical axis. However, for reasons that will soon become clear, economists prefer to represent inverse demand, that is, a diagram with the quantity on the horizontal axis and the price on the vertical axis. For a given price (on the vertical axis), the quantity consumed is measured on the horizontal axis. Demand decreases as the price increases; hence inverse demand curves are sloping downwards, as seen in Figure 32.

Since demand varies across hours of the year, we have numerous downward sloping inverse demand curves, one for each hour, as illustrated in Figure 32. For France, the left-most curve in Figure 32 corresponds to a summer morning, the right-most to a winter evening.

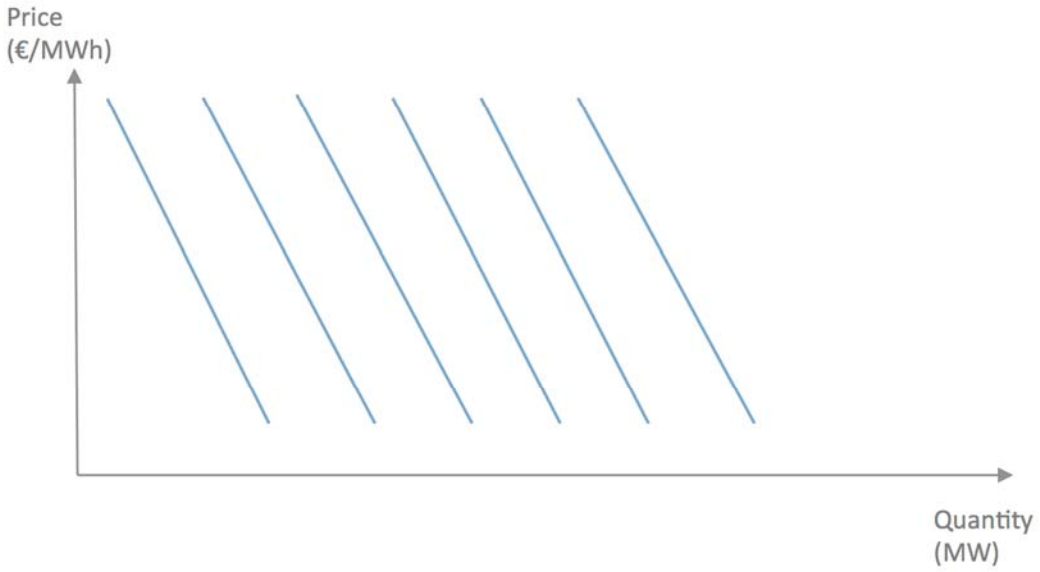

Figure 2: Inverse demand curves for different half-hours of the year. Source: Léautier, 2019, figure 2.2, page 19.

The basic model also assumes all customers are identical. Thus, they all have the same load duration curve and the same sensitivity to prices.

2.1.2 Supply

Variable and Fixed Cost of Production

The second assumption that we make initially is that only one generating technology is available.[7] This technology is characterized by a variable cost of production per unit, expressed in

€/MWh, and an hourly (amortised) fixed cost of production per unit, also expressed in €/MWh. The variable cost is essentially the cost of the fuel burned to generate electricity. It is assumed constant and denoted by c. It depends on the technology used and on fuel prices and usually ranges between 20 and 100 €/MWh for nuclear and fossil fuels-fired thermal plants. The variable cost is essentially zero for plants using intermittent renewable energy sources (RES) such as wind turbines and solar panels. Stored hydro is more complicated as the quantity of stored energy is limited and its utilisation (dispatch) is controlled by the system operator. While there is no direct cost to release the stored energy behind a dam, the relevant marginal cost is the opportunity cost of releasing water at a particular point in time rather than holding it to be released at other points in time. Calculating this shadow value of water is a complex stochastic dynamic programming problem.

The annual fixed cost of production includes the amortised capital cost for the technology (depreciation, return on investment, taxes, etc.) plus fixed operations and maintenance costs (O&M). The annual fixed cost is also assumed to be constant per unit of capacity, hence is sometimes called capacity cost, and is denoted by r. The fixed cost used in this discussion is the hourly fixed cost; that is, the annual fixed cost is divided by 8760 hours per year.[8] The magnitude of the fixed cost varies from one technology to another. As we will see later, a particularly relevant fixed cost for the models presented here is the fixed cost of an open cycle gas turbine (a peaking turbine).

In the example presented in this section, we use variable cost $c = 50$ €/MWh and annual fixed cost equal to 60 000 €/MW per year, hence $r = \frac{60{,}000\ €/MW}{8{,}760\ h} = 6.85$ €/MWh.[9]

Constant Returns to Scale in Power Generation

Consider a power plant of capacity K MW, which means it is impossible to produce output $Q > K$. As discussed above, the hourly capacity cost is assumed to be proportional to installed capacity and equal to rK. It must be paid every hour of the year. If the plant produces $Q \leq K$ MWh during a given hour, the total cost of production for this hour is $rK + cQ$.

This representation of the cost of producing electricity is, of course, an approximation. In reality, the variable cost increases as production gets closer to the maximum feasible capacity and the capacity cost per unit often decreases as capacity increases, as it includes a portion that is independent of capacity. For example, a power plant developer needs to pay lawyers to write up the contracts with the building contractors and equipment manufacturers. Planning, engineering, siting, regulatory and other legal fees are not proportional to the size of the asset; hence the power plant developer will pay an amount independent of the size of the asset. However, this approximation is close enough to reality that we can safely use it.

Under this approximation, electric power generation *for a given technology* exhibits constant returns to scale: producing 200 MWh using a 200 MW power plant costs exactly the same as producing them using two 100 MW plants *of the same technology*.

2.2 The Problem

Inverse demand is downward sloping and time-dependent. Furthermore, electricity cannot be stored economically on a large scale. This raises two questions: (i) how should we price electricity, and (ii) how much capacity should we build?

The term "should" is ambiguous. The problem is first solved from the perspective of a benevolent central planner; hence the questions can be rephrased as: 'what are the optimal electricity prices and generating capacity?' As often in economics, if competition is perfect, which is assumed in this chapter, the equilibrium reached by industry participants decentralises the optimum; hence the questions can be rephrased as: 'what electricity prices and capacity arise in equilibrium?'

Electrical engineers and economists have attempted to find a rigorous answer to these questions since the early days of the power industry in the 1890s. The formal answer was provided in 1949 by a young French economist, Marcel Boiteux, upon his return from World War II (Boiteux 1949 [1960], 1951; Dreze 1964; Turvey 1968).

Before describing the solution, observe that other goods share these two features, for example, hotel rooms and plane tickets. Neither can be stored: a seat on the 8 pm flight from New York to Paris must be "consumed" at 8 pm. Demand for both varies over time: more sea-side hotel rooms are requested in the summer than in the winter. The solution to the electricity pricing problem has been applied to these other industries, albeit with a major difference: retail price discrimination among consumers is added to peak-load pricing.

2.3 The Solution

2.3.1 Optimal Prices

A general result in economics is that to maximise the net surplus from consumption, price should be equal to the marginal cost of production and the marginal surplus from consumption. Understanding this result requires a few definitions.

Consumer Surplus

The consumer surplus, sometimes called the gross surplus or the surplus from consumption, is the surplus (or the utility, or the pleasure, or the benefit) that a representative consumer derives from consuming a given quantity of a good. To compute the gross surplus, economists estimate the surplus that a representative consumer derives from consuming each unit of the good, then add these surpluses.

Suppose that, for a given hour in a winter evening, a family consumes 5 kWh of electricity. One kWh goes to heating, and is valued at 30 cents, that is, generates a surplus of 30 cents. Another kWh goes to lighting, which matters slightly less, yielding a surplus of 15 cents. Another kWh goes to the various screens (television, computers, etc.), and is valued at ten cents. Finally, two other kWh go to domestic appliances (dishwasher and washer-drier) that could run later and are valued at five cents. The gross surplus from these five kWh is the sum of the surplus from each kWh: $30 + 15 + 10 + 2 \times 5 = 65$ cents.

Suppose now many infinitesimally small units of electricity are consumed. They can be ordered by decreasing per unit surplus: the first unit generates the highest per unit surplus, the next very slightly less, etc. If we plot the surplus per unit as a function of the number of units consumed, we obtain a downward sloping curve. This is the inverse demand curve, as presented in Figure 2.

The gross surplus derived from consuming quantity Q is represented by the hatched surface under the inverse demand curve on the left panel of Figure 3: it is the surface under the inverse demand curve up to the vertical line at quantity Q.

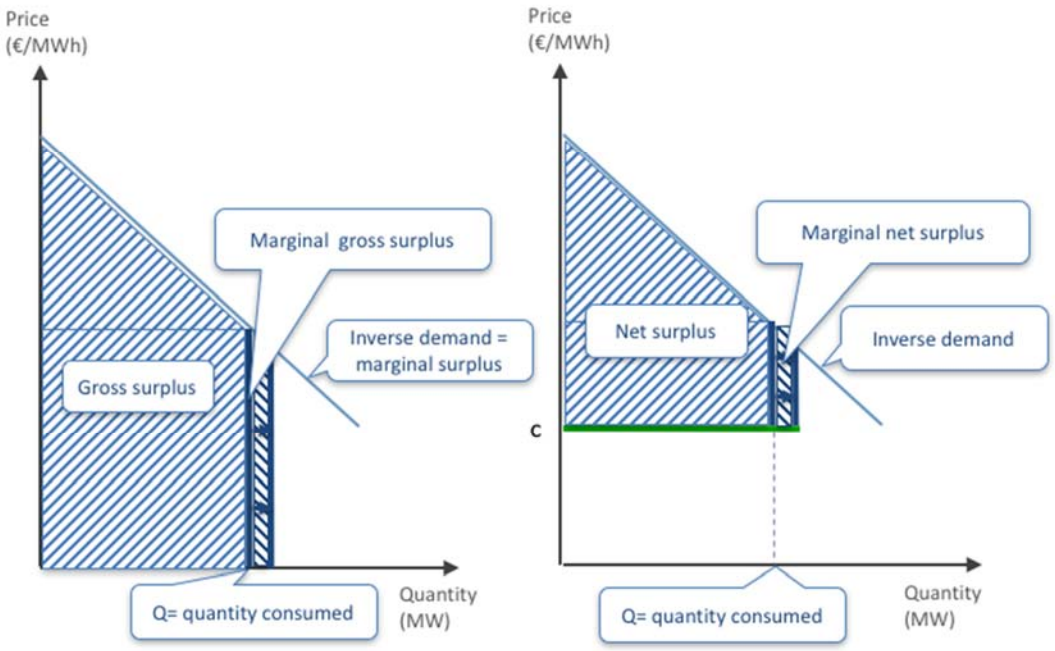

Figure 3: Gross consumer surplus and marginal gross consumer surplus (left panel) and net surplus and marginal net surplus (right panel). Source: Léautier 2019, figure 2.3, page 22.

Short-term Net Surplus

The *net surplus* from consumption is the surplus from consuming a given quantity minus the cost of producing this quantity. It measures the economic value generated by production and consumption. In the short term, the net surplus is the consumers' surplus minus the variable production cost. For example, if the variable cost of producing electricity is three cents/kWh, the short-term net surplus is $65 - 3 \times 5 = 50$ cents.

The net surplus from producing and consuming quantity Q is represented on the right panel of Figure 3 as the area under the inverse demand curve and above the production cost c, up to the vertical line at Q.

Marginal Surplus

The word "marginal" is often used in this chapter. It refers to the last unit produced or consumed, called the marginal unit, or to an attribute of the marginal unit. For example, the marginal surplus (sometimes called the marginal value) is the surplus of the last unit consumed. In the above example, the marginal surplus is equal to five cents/kWh. Similarly, the marginal cost is the cost of producing the last unit produced.

The marginal gross surplus is represented on the left panel of Figure 3, the marginal net surplus on the right panel. For every infinitesimally small quantity, the inverse demand is the marginal surplus.

Optimal Production and Consumption

The objective of a benevolent central planner is to maximise the net surplus. Production and consumption are short-term decisions; hence the short-term optimum is to maximise the short-term net surplus. The optimal production and consumption plan is: every unit which generates a positive short-term net surplus is produced and consumed, while no unit which generates negative short-term net surplus is produced. The optimal quantity produced and consumed therefore sets the marginal short-term net surplus to zero. Under reasonable conditions, this optimal quantity exists and is unique.

Equilibrium Price

Which price leads to the optimum? Consider it first from the perspective of the consumers. If the marginal surplus were higher than the price, consumers would consume more, since this would increase their surplus. Thus, they consume exactly all units whose surplus is higher than the price they pay. Consumption for any given price is such that the marginal surplus is equal to the price.

Consider now producers. If the price were higher than the cost, producers would produce more to capture positive profits. Production for any given price is such that the cost of the last unit produced (also called the marginal cost) is equal to the price.

Under reasonable conditions, there exists a unique equilibrium price such that supply is exactly equal to demand, that is, the quantity produced is exactly equal to the quantity consumed. At this price, the marginal surplus is equal to the price, which is also equal to the marginal cost. The marginal net surplus is equal to zero: the equilibrium leads to the optimum.

Off-peak and On-peak Prices

How does that insight apply to peak-load pricing? The key is to separate two different configurations: off-peak, that is when production is lower than installed capacity, producing a marginal megawatt-hour requires essentially only incremental fuel costs. Thus, the off-peak price is equal to the variable cost, which we have denoted as c. Consumption then adjusts to this variable cost, that is, consumption is such that the marginal surplus is exactly equal to the variable cost of production c. This situation is observed on the left of Figure 4.

On-peak, that is when production and consumption are precisely equal to installed capacity, the price is equal to the value of the last unit consumed, the value of the marginal megawatt-hour that fully utilises the available capacity. This situation is observed at the right of Figure 4.

Observe the duality between off- and on-peak. Off-peak, price is set by the variable generation cost and determines consumption. On the contrary, on-peak price is set by the value of the marginal megawatt-hour, that is, the megawatt-hour such that cumulative consumption equals capacity.

2.3.1.1 Supply Curve

A useful concept to examine markets and price is the supply curve that traces the short-run marginal cost, that is, the cost of producing a marginal unit of a good for various quantities of this good when capacity is already built. In our example, the supply curve is L-shaped.

Off-peak, the cost of producing one additional megawatt-hour is the variable production cost (essentially the fuel cost), denoted as c. The off-peak supply curve is the horizontal segment at the left of Figure 4. On-peak, when production equals installed capacity, the cost of producing one additional megawatt-hour exceeds the variable production cost, since this would require the deployment of additional capacity. The on-peak supply curve is then the vertical segment at the right of Figure 4.

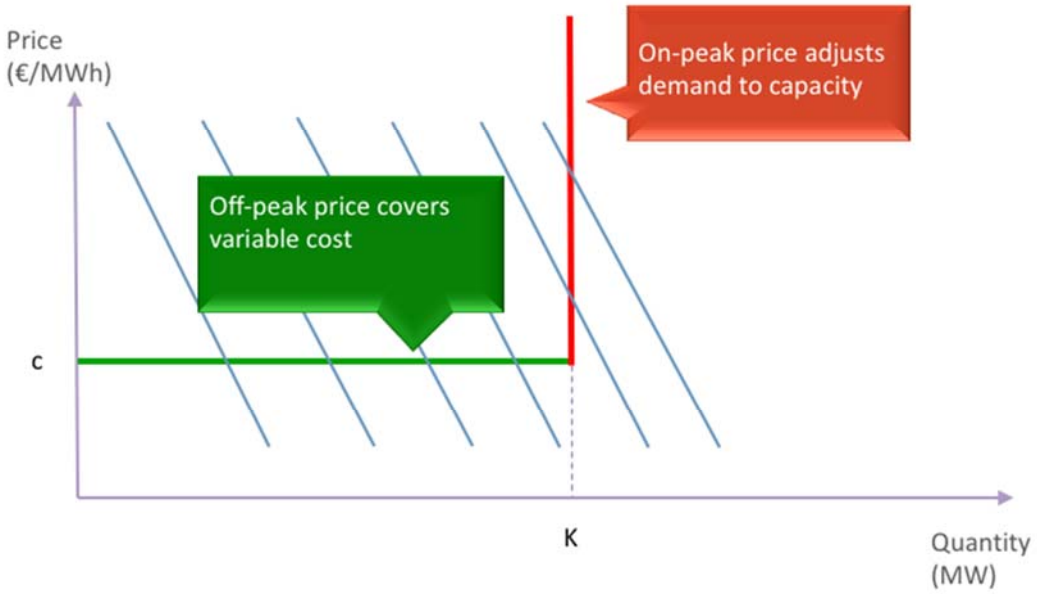

Figure 4: Demand curves and supply curve for a single generation technology. Source: Léautier 2019, figure 2.4, page 24.

2.3.2 Optimal and Long-term Equilibrium Capacity

Optimal Capacity

The capacity choice is a long-term investment decision; hence the long-term optimum is to maximise the average hourly long-term net surplus, which is the average hourly short-term net surplus minus the hourly capacity cost.

Consider adding a (marginal) megawatt of generation capacity. Installed capacity has no impact on surplus off-peak; hence the analysis is limited to on-peak hours. For every on-peak hour, since

consumption is exactly equal to installed capacity, adding one megawatt of generation capacity leads to the consumption of one additional megawatt-hour, which generates an hourly net surplus equal to the marginal surplus minus the variable cost of production c. For any on-peak hour, the left panel of Figure 5 presents the *short-term* net surplus, which is the area below the inverse demand curve and above the variable production cost c, and the marginal short-term net surplus from a marginal increment in generation capacity, which is the a rectangle of base the capacity increment and of height the marginal net surplus minus the variable production cost c.

The average hourly net surplus generated by a marginal megawatt of generation capacity is thus the average overall on-peak hours of these hourly net surpluses. It is represented by the hatched triangle in Figure 6.

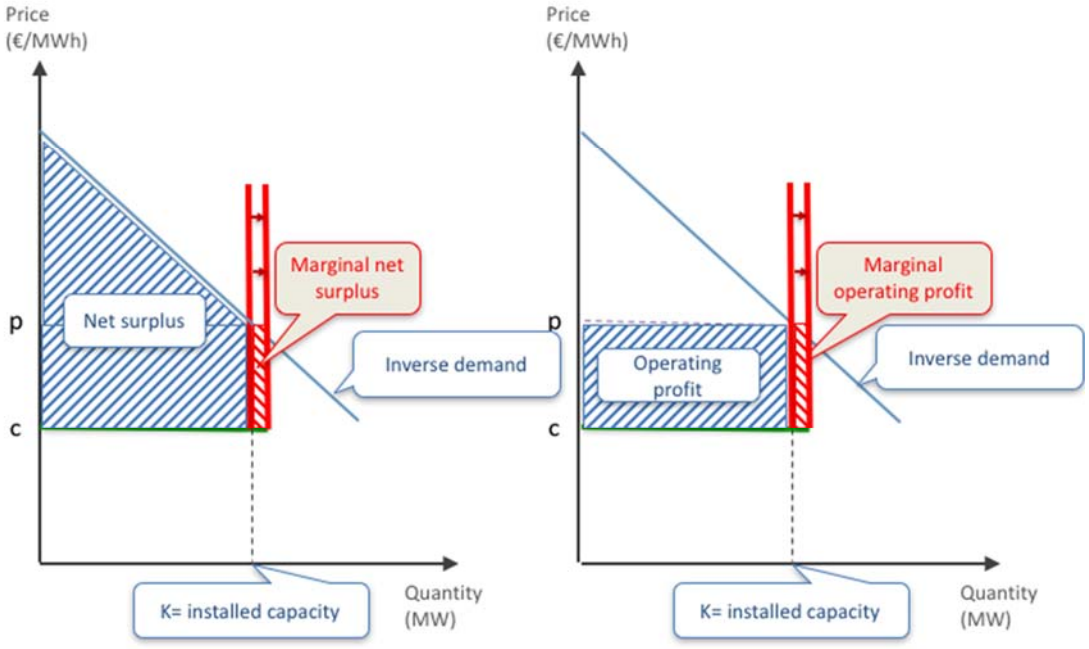

Figure 5: Short-term net surplus and marginal short-term net surplus (left panel), and operating profit and marginal operating profit (right panel) for any on-peak hour. Source: Léautier 2019, figure 2.5, page 25.

On the other hand, adding one megawatt of generation capacity costs the hourly capacity cost r. Under reasonable conditions, there exists a unique optimal capacity which precisely equates the average hourly marginal net surplus and the hourly capacity cost.

Long-term Equilibrium Capacity

The equality between average hourly marginal net surplus and hourly capacity cost is both an optimality condition, that is, it maximises the net surplus, and also an equilibrium condition. Consider a producer adding a (marginal) megawatt of generating capacity. She realises no

operating profit off-peak since she sells at a price equal to her variable cost of production c; hence her analysis is limited to on-peak hours. For any on-peak hour, the right panel of Figure 5 presents her operating profit, which is a rectangle of base the generating capacity and of height the price p minus the variable production cost c, and the marginal operating profit from a marginal increment in generating capacity, which is the rectangle of base the capacity increment and of height the price p minus the variable production cost c. A comparison of the left and right panels of Figure 5 shows that the marginal operating profit is exactly equal to the marginal net surplus, since the price is equal to the marginal surplus, even though the net surplus exceeds the operating profit.

If competition is perfect, producers build capacity until the last unit precisely breaks even, that is, until the average hourly marginal operating profit is precisely equal to the hourly fixed cost r. Otherwise, if the average hourly marginal operating profit exceeds (resp. was lower than) the hourly fixed cost, producers would increase (resp. decrease) installed capacity. This is known as a "free entry condition".

Since marginal operating profit is equal to the marginal net surplus, the long-term equilibrium is also the optimum.

Resulting Price Structure

The above story seems simple enough. However, it has profound implications. The price in off-peak hours is the variable cost, which we assumed here to be around 50 €/MWh. Meanwhile, the on-peak margin has to cover the capacity cost, which we have assumed to be 60 000 €/MW/year. The price structure is illustrated in Figure 6.

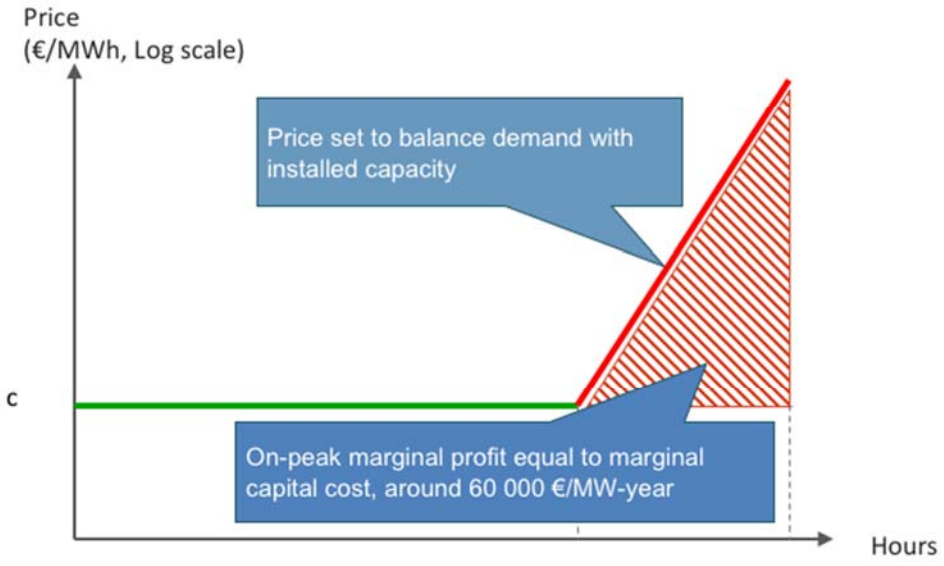

Figure 6: Price structure. Hours are represented on the horizontal axis, prices on the vertical axis. For most hours, price is equal to the variable cost c. On-peak, price rises above the variable cost. The operating margin is equal to the fixed cost. Source: Léautier 2019, figure 2.6, page 26.

Different combinations of on-peak hours and on-peak prices could yield the same on-peak margin: many on-peak hours at low on-peak prices, or few on-peak hours at high on-peak prices. Which combination occurs depends on the sensitivity of power demand to prices. If electric power demand is not very sensitive to prices, which is currently the case, the on-peak price will rise very high; hence on-peak hours will be few. On the other hand, if electric power demand were very sensitive to power prices, the equilibrium price would not rise very high; hence the number of on-peak hours would be large.

Demand sensitivity to price is determined by two factors: (i) the share of consumers adapting their consumption in response to spot wholesale prices, and (ii) each consumer's sensitivity to spot wholesale price.

Suppose first that all consumers are identical and respond to wholesale spot prices. Léautier (2019, pp. 82-83) estimates that, if individual customers' sensitivity is at the high end of empirical estimates, on-peak would last around 600 hours per year and price would rise to around 350 €/MWh, seven times greater than the off-peak price.

On the other hand, if all customers respond to price, but their individual sensitivity is at the low end of empirical estimates, the on-peak period would last less than 150 hours per year and price would rise to around 1200 €/MWh, more than 20 times the off-peak price.

In reality, not all customers respond to prices, nor are they identical in other dimensions. Suppose for example 20 per cent of demand responds to wholesale prices, which is higher than in most markets today, but a reasonable target in a few years once demand response policies are implemented (advanced or "smart" meters that can measure consumption over very short periods of time and spot prices that vary accordingly).[10] Léautier (2019) estimates that, if individual customers' sensitivity is high, on-peak would last around 200 hours per year and price would rise to around 800 €/MWh, more than 15 times the off-peak price. If individual customers' sensitivity is low, on-peak would last around 50 hours per year and price would rise to around 3300 €/MWh, more than 60 times the off-peak price.

Basic microeconomics suggests that, due to low demand sensitivity to prices, on-peak prices 20 or 60 times higher than the off-peak prices are to be expected and are perfectly legitimate.

Long-run Marginal Cost

In the long run, before capacity is built, producing one megawatt-hour requires building one megawatt of capacity and burning fuel for one hour. The hourly long-run marginal cost is thus $(c + r)$.

The free entry condition implies that the *average* hourly price is equal to the hourly long-run marginal cost. An additional megawatt of capacity produces electricity in every hour; hence its average hourly revenue is the average hourly price. If the latter exceeded (resp. was below) the hourly long-run marginal cost, producers would profitably increase (resp. decrease) installed capacity.

3 A More Realistic Story

Reality appears to be much more complex than the above example. Surely, the economics of large and complex power systems cannot boil down to such a simple story. Well, in fact, it (almost) can, even though multiple new elements must be added: (i) demand does not fully respond to spot wholesale prices; (ii) there exists more than one technology to produce power; (iii) electric power is transported across continents over transmission networks that have capacity constraints which may limit transfers of power from one location (node) on the network to another, thereby causing congestion and yielding optimal wholesale prices that vary from one location to another on the network; and (iv) running a power system requires energy, but also operating reserves and other ancillary services to accommodate uncertain variations in demand and generating plant outages. This section examines each consideration in turn and concludes that adding them does not significantly alter the basic peak-load pricing story. Finally, this section discusses "security of supply", a term loosely used in policy debates, which covers, in fact, three distinct time horizons.

The main message of this section is that the standard model does surprisingly well at explaining the main economic intuitions about the attributes of competitive wholesale power markets. It

misses key ingredients, such as inter-temporal linkages and geographical differentiation, hence some numbers may not be completely accurate, but its logic is robust.

3.1 Non-price-responsive Demand

The standard model assumes that all consumers (i) are identical and that (ii) they adjust their consumption to respond to variations in the spot price of electric power. The first assumption is clearly not realistic: customers have different uses for power, hence different needs and valuations. At best, we can group customers by classes (for example, industrial, commercial, residential). However, this richness in usage across customers does not modify the structure of the analysis: all we need is a downward sloping aggregate demand curve, which can be built by the aggregation of different customers' demand curves.

3.1.1 Retail and Wholesale Prices

The second assumption is not met in reality either. Most customers purchase electricity from a "retailer" (or "distributor" or "supplier" or more generally a "load-serving entity"), usually through contracts of varying durations, while retailers purchase power from producers on wholesale markets also pursuant to contracts of varying durations.[11]

This general description covers multiple situations. In Europe, wholesale markets are decentralised, that is, buyers and sellers enter into bilateral transactions, that the market operator aggregates and ultimately adjusts to meet network feasibility constraints. In North America, wholesale markets are centralised, that is, the market operator runs an auction to collect all offers from producers and demand from retailers and consumers, and determines the equilibrium production and price. This distinction is ignored in this text since both market organisations should lead to the same outcome under perfect competition.

Wholesale markets exist for multiple dates. The most important is the spot market since the hourly wholesale spot price defines the value of energy at every hour. In most markets, the spot market is, in fact, a day-ahead market, not a true real-time market. Since technological constraints imply that (most) power plants must decide today whether or not to be online tomorrow, buyers and sellers agree today on the quantities each will buy and sell, and on the price for electricity for tomorrow between, for example, 4 and 5 pm. Since demand and supply conditions may vary between 4 pm today and tomorrow, adjustment markets also exist, which also produce prices for electricity for tomorrow between 4 and 5 pm. Thus, most electric power markets are "two-settlement" markets, in which the price for electricity for a given hour is settled twice, once day-ahead and then in the day-of adjustment market. This introductory chapter does not open the "black box" of complexities associated with wholesale price formation in practice and assumes a single wholesale spot price exists for a particular hour. For discussions of the details of wholesale power market designs in several different countries see Chapters 3, and 5-10 in this Handbook.

Forward wholesale markets also exist, where producers and buyers can exchange power for the next weeks, months and years. In addition, financial instruments are available to hedge future prices. For example, a producer can sell a future's contract, which pays the difference between the spot price at a given date and the forward price.

The structure described above applies to most commodities, for example, oil, metals, agricultural products, etc. In most of these industries, customers face wholesale spot prices, sometimes with a

lag. For example, drivers pay the wholesale spot price for gasoline at the pump (plus a retail margin and a variety of taxes) and wheat retail prices follow the wholesale spot prices.

In the power industry, by contrast, most customers pay a constant retail price, sometimes called a "flat rate", that does not vary from hour to hour. Historically, meters could only record consumption between two readings separated by 30 to 90 days; hence customers were paying the same price for every megawatt-hour they consumed, irrespective of the true value of energy at the hour of consumption or the marginal cost of supplying it.

If all customers face a constant retail price, the inverse demand is a vertical line. This may have been a reasonable representation of demand twenty or thirty years ago, but is now unrealistic in most markets. Today, most electro-intensive customers purchase directly from wholesale markets or from intermediaries that convey them variable wholesale market prices.[12] Smart meters and enhanced communications technology enable retailers to record hourly demand and to charge a different price for every hour, even for residential customers, although most retail contracts, in particular for residential customers, still offer a flat rate.

As long as a positive fraction of customers responds to the spot price, the aggregate demand remains downward sloping.

3.1.2 System Operator and Power Exchange

At this juncture, it is useful to briefly discuss the primary organisational attributes of wholesale power markets in more detail. Power markets require a system operator (SO) to physically control supply and demand in real-time. More detailed discussions of these organisational arrangements and their variations across wholesale markets can be found in Chapters 3 and 5-10 of this Handbook. This feature is unique to the power industry. In most other markets, physical delivery of the underlying commodity is decentralised: no single central entity controls the production of all oil fields, the consumption of all oil refineries and the movement of all oil tankers; rather every market participant optimises the physical movements of its own assets.

In the power industry, the SO has her fingers on the switch(es): in real-time, she is allowed to turn power plats on or off and to curtail customers. This function is conceptually different from the power exchange (PX), which organises the wholesale market(s) in Europe, that is, provides a platform for producers and consumers to sell and buy energy. In the US, the functions of the PX and the SO are integrated as discussed below.

A simple example illustrates the articulation between PX and SO. Consider a producer who sells 100 MW for the next day from 4 pm to 5 pm into the PX. The next day at 4:30 pm, the producer's plant suddenly trips and cannot produce at all. To balance the system, that is, to ensure that supply meets demand, the SO must turn on another plant. She then requests the defaulting producer to pay for this additional energy.

In most markets in the United States, the SO is also the PX. The underlying argument for this choice is that engineering constraints perfectly understood and mastered by the SO also structure the work of the PX. On the contrary, in most European markets, the SO and the PX are different and, in fact, multiple PXs exist for a single SO and vice versa. This structure reflects the implicit choice of power markets' designers to place less weight on engineering and physical constraints. In the end, it is not clear whether the results of these two institutional designs are very different.

In a world that satisfies all of the assumptions of perfect competition the results should be the same.

3.1.3 Administrative Curtailment and the Value of Lost Load

When demand is vertical, the peak-load pricing logic applies differently: demand can no longer be adjusted to capacity through an increase in price. When demand (at any price) exceeds capacity, administrative curtailment is required. The SO implements rolling blackouts to balance supply and demand, that is, selectively shut down parts of the power system for a few hours. The government usually approves the curtailment plan.[13]

When demand is curtailed, the SO should value electricity at the Value of Lost Load (VOLL) to define an efficient curtailment plan (Schroder and Kuckhsinrichs 2015). Unless rationing is efficient, the (marginal) VOLL, that is, the value a user is willing to pay for a (marginal) megawatt-hour when rationing occurs, is higher than the (marginal) value of power when it does not. An example illustrates the argument.

Consider two customers: customer 1 uses electricity for heating with value 200 €/MWh, while customer 2 uses electricity for lighting with value 100 €/MWh. When there is no rationing, each consumes one kWh and the marginal value for the system is 100 €/MWh.

Suppose now rationing must be implemented and only 1.9 kWh is available. If rationing is efficient, the lowest value usage is curtailed: customer 1 uses 1 kWh to heat her house, and customer 2 is rationed and uses 0.9 kWh to light her house. How much would the SO value the marginal 0.1 kWh? He would deliver it to user 2, hence value it at 100 €/MWh. Thus, the marginal values with and without rationing are equal.

In practice, however, rationing is often inefficient, that is, the SO cannot curtail consumer 2 alone and instead must curtail both. Each will have 0.95 kWh available for heating and 0.95 kWh available for lighting. A marginal 0.1 kWh would be used for heating and lighting, hence would be valued at 150 €/MWh, which is higher than without rationing.

There are many reasons for inefficient rationing. In many cases the SO cannot curtail individual consumers, but only groups of consumers on a controllable distribution circuit (although in the not-too-distant future, smart meters with two-way communications and advanced monitoring and control of distribution circuits will facilitate rationing of individual consumers). Furthermore, the VoLL is uncertain both in the aggregate and across individual consumers. We turn to this issue next.

3.1.4 What is the VoLL?

In the above two-usage example, the SO is able to compute the VoLL for each customer. This is not the case in reality since a customer's VoLL depends on a number of factors. First, it depends on usage. Students in a classroom hit by a power outage are probably not willing to pay much to get the light back and resume the course. In fact, most would be willing to pay (at least a small sum) to enjoy a break in the sunny courtyard. On the other hand, a patient receiving open-heart surgery would be willing to pay a significant sum to avoid curtailment.

Second, the VoLL depends on the duration of the outage. Supermarkets have deep-frost fridges, which keep perishable products at extremely low temperatures. They are not willing to pay much

to avoid an outage that lasts only a few minutes. In fact, in some instances, they are willing to reduce their fridge's consumption of electricity and resell part of their energy into the market for a short duration. On the other hand, they would be willing to pay a large amount to avoid an outage lasting a few hours, which would destroy all their stocks.

Third, the VoLL depends on the information given to customers. If you know you will be curtailed tomorrow at 10 am, you do not step in an elevator at 9:59 am. On the other hand, if the curtailment catches you unaware and you end up stuck in a cramped elevator with foul-smelling colleagues, you will be willing to pay a significant sum to get power back hence terminate your ordeal.

It is therefore not surprising that estimates of the VoLL vary in an extremely wide range, from 2000 £/MWh in the British Pool in the 1990s to 200 000 $/MWh (Cramton and Lien, 2000 : Cramton P. and J. Lien, "Value of Lost Load", Mimeo, University of Maryland). Schröder and Kuckshinrichs (2015) provide a recent survey. As an illustration, we use in this chapter 20 000 €/MWh as an estimate of the VoLL, which is consistent with the security of supply standard used in France.

3.1.5 Resulting Demand Curves

When all customers face a constant retail price, the inverse demand curve is vertical for prices up until the VoLL and horizontal afterwards, unless there are administrative price caps below VoLL (Figure 7).[14] When only a small fraction of customers respond to spot prices, there may also be instances when administrative curtailment is required. In that case, inverse demand is a steeply sloping line up until the price equals VoLL, and then a horizontal line when the price is equal to the VoLL (Figure 8).

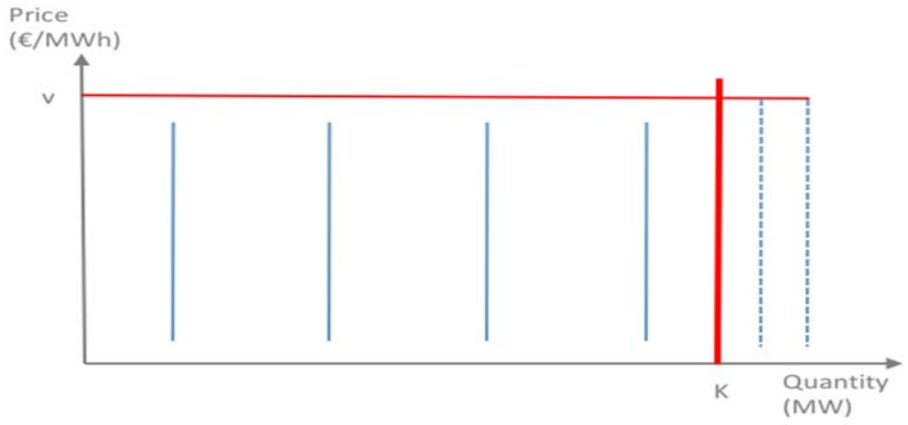

Figure 7: Demand curves if demand is perfectly inelastic: in every hour, demand does not vary with price. When demand is lower than installed capacity K, it can be entirely served. When demand exceeds installed capacity K, it must be reduced through involuntary curtailment. The

price is then set at the VoLL: inverse demand is a horizontal line at the VoLL. Source: Léautier, 2019, figure 2.7, page 33.

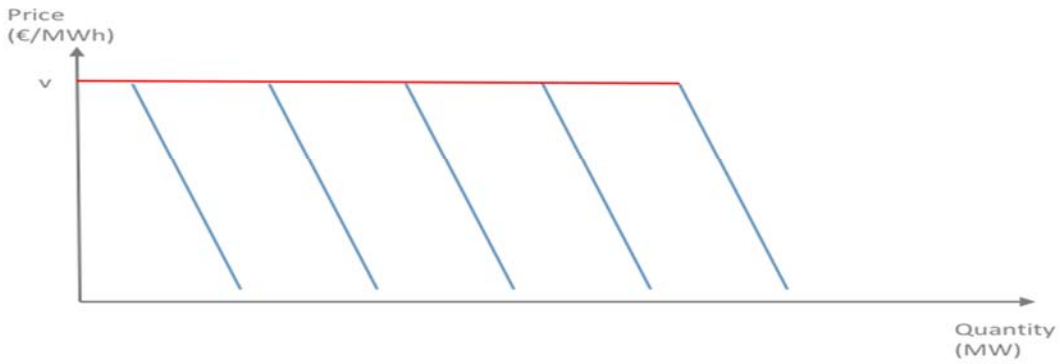

Figure 8: Demand curves if demand is partially elastic. Demand is slightly reduced as price increases, up to the VoLL. Source: Léautier, 2019, figure 2.7, page 33.

3.1.6 Resulting Price Structure

The peak-load pricing logic still applies to that inverse demand curve. The only difference is that the visible hand of an administrative intervention replaces the invisible hand of market forces to adjust demand to available capacity through curtailment and to set the wholesale price at VoLL.

When demand does not respond to price, the latter is equal to the variable cost (around 50 €/MWh in our illustrative example) for almost all hours and is set by the SO at the VoLL for the remaining hours, during which power is curtailed. This is illustrated in Figure 9.

When demand partially responds to price, the latter is equal to the variable cost for most of the hours. When demand (for price equals variable cost) is equal to installed capacity, price increases and price-responsive customers reduce their demand accordingly. As was the case in Section 2, this is voluntary demand reduction and not involuntary curtailment. As long as the price is lower than the VoLL, the SO does not curtail any customers administratively: responsive demand balances supply and demand as prices rise.

If the price rises up to the VoLL, the SO starts curtailing non-price-responsive customers and sets the price at the VoLL, which represents the value of an additional megawatt-hour. Price-responsive customers adapt their demand to this price. This price structure is illustrated on Figure 10.

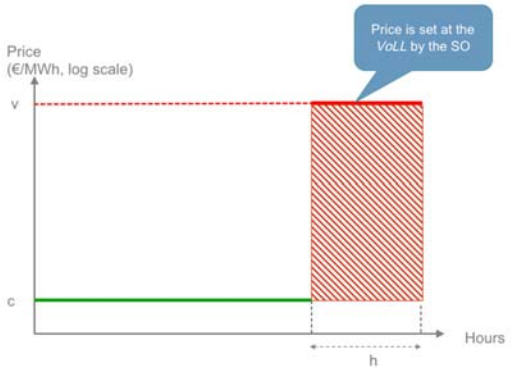

Figure 9: Price structure when no customer is price-responsive. The SO starts curtailing customers when demand (at the variable cost of production) is equal to installed capacity and sets the price to VoLL. Source: Léautier, 2019, figure 2.8, page 36.

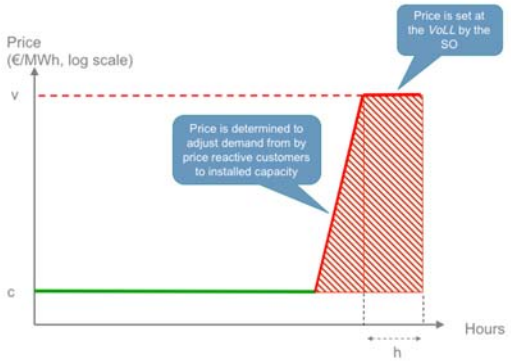

Figure 10: Price structure when a fraction of customers is price responsive and curtailment occurs. The price rises on-peak as in Section 1. If the price rises to the VoLL, the SO starts curtailing constant price customers and sets the price at VoLL. Source: Léautier, 2019, figure 2.8, page 36.

3.1.7 When is Administrative Curtailment not Required?

Figure 9 represents the pricing structure of the past when no customer was price responsive. Figure 10 represents the evolving pricing structure when a small – but growing – fraction of customers is price responsive. In a few years, the fraction of price responsive customers will be large enough that the pricing structure is that of Section 2. This naturally raises the question: 'How much demand response is required for curtailment never to be necessary?'

Intuitively, the higher the share of demand responding to the wholesale spot price, the lower the probability administrative curtailment is required. Numerical simulations presented in Léautier

(2019) put a number on this intuition. Rationing is not required for the optimal generation mix when more than 3.9 per cent of demand is price-responsive if the price elasticity of demand is low. If demand elasticity is high, rationing is no longer required when more than 13.9 per cent of demand is price-responsive. *This result may seem counter-intuitive*: a less elastic demand results in less curtailment. The intuition is that, for a given share of price responsive customers, optimal capacity is *higher* when demand is less elastic: since customers reduce their demand less when price raises, it is optimal to increase installed capacity. Since capacity is higher, curtailment is less frequent.

These results are obtained under specific assumptions on the shape of the demand function and the nature of the uncertainty. Further research is required to refine these values. Still, they are good news. If less than 20 per cent of demand responding to price is indeed sufficient for curtailment never to occur at the optimal capacity, rationing will soon be a practice of the past.

3.1.8 Optimal Capacity

As in Section 2, the optimal capacity is such that the average hourly marginal operating profit of a megawatt of capacity is equal to its hourly fixed cost (Boiteux 1956; Turvey 1968).

If no customer responds to price, the marginal operating profit is VoLL minus production cost, usually approximated by VoLL, times the number of curtailment hours. This corresponds to the hatched rectangle in Figure 9. Suppose for example the system operator sets the VoLL 20 000 €/MWh. Thus, if the marginal fixed cost of capacity is assumed to be 60 000 €/MW/year, the optimal capacity is such that power is curtailed three hours per year on average since 60 000 €/MW/year = 20 000 €/MW/hour times three hours/year.

If a fraction of customer responds to price, the marginal operating profit is the price minus the production cost, where the price is set to balance demand from price responsive customers with installed capacity (hatched triangle in Figure 10) then set at the VoLL by the SO (hatched rectangle in Figure 10).

3.1.9 Engineering Generation Adequacy Criterion

Power engineers do not design power systems using the VoLL. Rather they use a physical generation adequacy criterion, for example, ‘available generation should exceed demand for all but three hours per year on average’. The criterion is determined administratively and may or may not coincide with the economic optimum.

To meet the criterion specified in the above example, engineers and statisticians first compute the distribution of possible future peak demands, considering different scenarios for weather and economic growth (and other relevant variables), and determine the $(8{,}760 - 3)/8{,}760 = 99.966$ percentile of the distribution. Suppose, for example, it is 108 GW. The probability that demand exceeds 108 GW is only 0.034 per cent, which is equivalent to say that, on average, demand will be lower than 108 GW for 99.966 per cent of the time or all but three hours per year.

Then, system planners decide how much generation capacity should be built, so that, on average, generation assets can produce 108 GW for a few peak hours. To do so, planners take into account unplanned outages in power generation units (sometimes called forced outages). For example, suppose they assume a forced outage rate of 7 per cent, which implies that, on average, generation units produce 93% of their nominal capacity durning peak hours. Therefore, the

adequate generation capacity, which guarantees that, on average, demand is met for all but three hours per year, is 108/0.93 = 116 GW.

This does not mean that three hours of rolling blackouts will occur every single year. If the weather is mild and plants' operating conditions are good, no rolling blackout may occur for one or more years. On the other hand, if the weather is unfavourable (a winter colder than average in Europe or a summer hotter than average in the United States), and if plants' operating conditions are weak, three or more hours of rolling blackouts may occur. On average, however, if the engineers and statisticians' computations are correct, rolling blackouts should occur about three hours per year.

The outcome of the generation adequacy computation is often expressed as a capacity or reserve margin, that is, the generation capacity minus the expected peak demand, as a fraction of expected peak demand. In the previous example, suppose, for example, that the expected peak demand is 100 GW. The capacity margin is $\frac{116-100}{100} = 16\%$. Adequate generation capacity is 16 per cent higher than expected peak demand: 8 per cent is due to demand being higher than its expectation and another 8 per cent is due to forced outages in production. This number is actually representative of capacity margins used by engineers up until renewable energy sources were introduced. If the capacity margin exceeds 20 per cent, too few on-peak hours occur, wholesale spot price remains close to marginal cost and capital cost cannot be recovered. If the capacity margin falls below 10 per cent, rolling blackouts are likely to exceed three hours per year on average.

The capacity margin loses its meaning as RES enter electricity markets since they produce on average 15-50 per cent of the time (solar at the low end and off-shore wind in good locations at the high end of this range) and their production cannot be controlled. That is, production is intermittent and responds to variations in sun and wind availability: unlike conventional dispatchable generation it cannot be dispatched economically by the system operator.[15]

While still widely in use today, traditional generation adequacy criteria should become almost irrelevant in the future as demand becomes fully price responsive. As demand becomes progressively more price responsive, it will adjust to available supply through an increase in price, not through administrative curtailment as long as investment in generating capacity reflects this reality as well. Anticipating this trend, some countries such as New Zealand have abolished the engineering reliability criterion altogether.[16] Other countries are holding on to the criterion. In the US, most regions continue to rely on traditional resource adequacy criteria. However, ERCOT (Texas) does not officially have a resource adequacy criterion either. ERCOT has an "energy-only" market. It has constructed an operating reserve demand curve (ORDC) that allows for "scarcity pricing" when generation supplies get tight. The ORDC reflects an estimate of VoLL, demand variation and associated uncertainty, generating unit outage rates and loss of load probabilities. See Chapter 7 of this Handbook for a detailed discussion of the ERCOT market.

3.1.10 Formal Equivalence between the VoLL and the Generation Adequacy Criterion

Both approaches are formally equivalent. Since the product of the (expected) number of hours of curtailment times the VoLL is equal to the marginal fixed cost of capacity, choosing a generation adequacy criterion is mathematically equivalent to choosing a VoLL and vice versa. In our example, if the capital cost of a peaking turbine is 60 000 €/MW/year, a VoLL set at 20 000

€/MWh is mathematically equivalent to a generation adequacy criterion that 'available generation should exceed demand for all but three hours per year on average'. The higher the generation adequacy criterion (that is, the lower the expected number of curtailment hours), the higher the VoLL.

However, setting a VoLL or setting a generation adequacy criterion leads to dramatically different market designs. In the first approach, known as the "energy-only" market design, regulatory intervention is in theory limited to setting the price for electricity when rolling blackouts are required and designing a plan for curtailments under extreme conditions.[17] In the second approach, policymakers require the SO to set up an additional "capacity mechanism" to guarantee the generation adequacy criterion is met, which is a much more complex market design.

3.1.10.1 The No Rationing Puzzle

The economic analysis presented above suggests that, if demand is not very elastic, rationing should occur for a few hours per year to cover the fixed cost of generation. Suppose, for example, demand is perfectly inelastic. As seen previously, if rationing never occurs, the wholesale spot price is always equal to the variable cost of production and no fixed cost recovery occurs.

In practice, in developed countries with robust electric power systems, there are very few hours when curtailments take place due to inadequate supplies of generation, if any.

Multiple reasons explain this. First, in the US, as previously mentioned, there is really a set of "emergency responses", for example, allowing operating reserves to fall below targets, voltage reductions, emergency payments to customers and on-site emergency generators to reduce demand or provide energy to the network, etc. While economic analysis suggests that the price should rise to VoLL (or towards VoLL*)* during these conditions, it is not always so. Second, the mandated "reserve margins" are probably too high. In other words, policymakers and system operators attempt to limit price increases and customer rationing, while economists argue they are necessary (when demand is inelastic) to cover fixed costs of generation.

Policymakers and system operators have to choose between either (i) rationing customers for a few hours or (ii) failing to cover fixed generation costs from energy market revenues.

This difficult choice arose with the restructuring of the power industry. When generation was a regulated regional monopoly, policymakers and the utility could agree on a sufficiently high generation adequacy criterion that rationing never occurred under reasonable scenarios or equivalently that rationing occurred only under scenarios so severe they were politically acceptable (for example, the worst blizzard of the century). The resulting excess generation capacity was then included in the regulated asset base.

This practice was one of the first casualties of the restructuring of the power industry. By construction, unregulated generating assets cannot be included in the regulated asset base. Interestingly enough, this was not a concern at the onset of restructuring, since most restructured power systems had excess generation capacity (which partially explains why they were restructured in the first place).

This choice stands at the heart of market design. On the one hand, policymakers may opt to recreate the previous situation, by mandating a high generation adequacy criterion and paying

quasi-regulated payments to the owners of generating assets to invest and maintain the required generating capacity. These are called "capacity mechanisms" and are reviewed in Chapters 5-10 in the context of individual market designs discussed there. On the other hand, policymakers may recognise that one of the main objectives of the restructuring of the power industry was to have market participants and not bureaucrats make investment decisions and take the accompanying risks. They will then accept that the price rises to the VoLL (hence a fraction of customers is rationed) for a few hours, in the hope that these high prices will spur enough demand response that rationing is no longer required.

Policymakers in the US, with the notable exception of ERCOT (most of Texas), have chosen the "capacity mechanism" route.[18] In Europe, as of this writing, most countries have a capacity mechanism or are designing one.

Further discussion of capacity mechanisms applied in different markets is offered in Chapters 5-10 of this Handbook. See also Cramton and Stoft (2005), Joskow and Tirole (2007), Joskow (2008), Léautier (2016), Léautier (2019) Chapter 9 and Keppler (2017). From this chapter, readers should simply take away that setting up a capacity mechanism is really a legitimate political choice, not an economic imperative.

3.2 Multiple Technologies

These price and investment concepts extend naturally to the more realistic case of multiple generating technologies. When only one technology is present, we have seen that the supply curve is L-shaped: horizontal at the variable production cost until capacity, then vertical when capacity is reached. When multiple technologies are present, the supply curve is a staircase, that is, a succession of Ls. (Boiteux 1956; Turvey 1968).

3.2.1 Characteristics of Multiple Technologies

To make things concrete, suppose three technologies are available: nuclear, combined cycle gas turbine (CCGT) and open cycle gas turbine (OCGT). Table 1 presents illustrative estimates of the variable and fixed costs of each technology.

	Fixed cost (€/MW/year)	Fixed cost (€/MWh)	Variable cost (€/MWh)
Nuclear	299 000	34	10
Combined Cycle (CCGT)	72 000	8	90
Gas Turbine (OCGT)	53 000	7	130

Table 1: Illustrative fixed and variable cost of different production technologies. Source: International Energy Agency (2010: International Energy Agency, 2010. Projected costs of generating electricity. OECD/IEA) median case with two modifications: gas price 40 €/MWh and CO_2 price 50 €/ton.

As shown in Table 1, the technologies are ordered by increasing operating cost: nuclear is cheaper than CCGT, which is cheaper than OCGT. The technologies are also ordered by decreasing fixed costs: if a technology has lower short-term marginal cost than another, it has higher fixed costs. This makes sense: if a technology was cheaper to build and to run than all

others, it would be the only one installed. Similarly, no one would install a technology more expensive to run and build than the others.

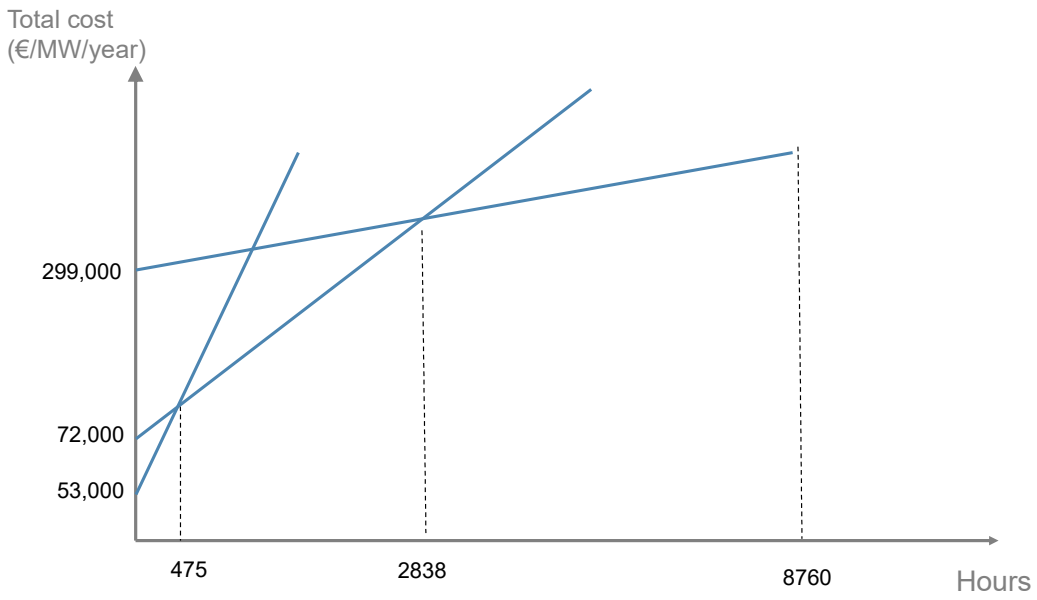

Figure 11: Screening curves for three generation technologies: nuclear, CCGT and OCGT. Hours are represented on the x-axis and the y-axis presents the total annual cost of producing one MW during a strip of any number of hours. Source: Léautier, 2019, figure 2.9, page 41.

3.2.2 Screening Curves and Optimal Usage of Different Technologies

The trade-off fixed versus variable cost of generation is illustrated on the screening curves, presented in Figure 11. Hours are represented on the *x*-axis and the *y*-axis presents the total annual cost of producing one MW during a strip of any number of hours. For example, producing one MW for one hour using nuclear technology costs €299 000 of fixed cost, plus €10 per hour of production. The total cost of serving a strip of hours using a nuclear power plant is, therefore, a straight line of intercept €299 000 and slope €10 per hour. Similarly, the total cost of serving a strip of hours using a CCGT is a straight line of intercept €92 000 and slope €70 per hour.

Figure 11 illustrates that hourly long-run marginal costs are decreasing along with operating costs. Consider the OCGT. Its fixed cost is lower than that of the CCGT. Its variable cost has to be high enough so that the sum of fixed and variable costs crosses the total cost of the CCGT; otherwise the CCGT would never be installed. The same argument shows that the long-run marginal cost of the nuclear technology is lower than that of the CCGT.

This particular cost structure translates into a differentiated usage pattern. Remember that demand varies significantly across months, weeks, days and hours of the day. The issue is: under which circumstances should a specific technology be turned on? Since nuclear is the most expensive to build and the cheapest to run, it should run all the time. It is the "baseload" technology. In markets where nuclear is not present, coal is often the baseload technology.

At the other extreme, since the OCGT is cheap to build and expensive to run, it should be turned on for high demand situations (winter evenings in Europe, summer afternoon in the United States). This corresponds to a peaking usage. Finally, CCGT being the intermediary technology (sometimes called semi-base or mid-merit or load-following), it starts running for intermediary demands, for example, 2 000-5 000 hours per year.

These results can be illustrated using the screening curves presented in Figure 12 and Figure 13. The three lines cross: the total cost of CCGT crosses the total cost of OCGT at 475 hours of utilisation and crosses the total cost of nuclear at 2 838 hours of utilisation. As presented in Figure 12 to produce a strip of hours lasting more than 2 838 hours, nuclear is the cheapest technology, measured in € per MW per year. Figure 13 shows that to produce a strip of hours lasting less than 475 hours, OCGT is the cheapest technology. For an intermediate number of hours, CCGT is the cheapest technology. This analysis constitutes an excellent and highly illustrative "first approximation". It is rigorously exact if and only if no customer responds to prices.

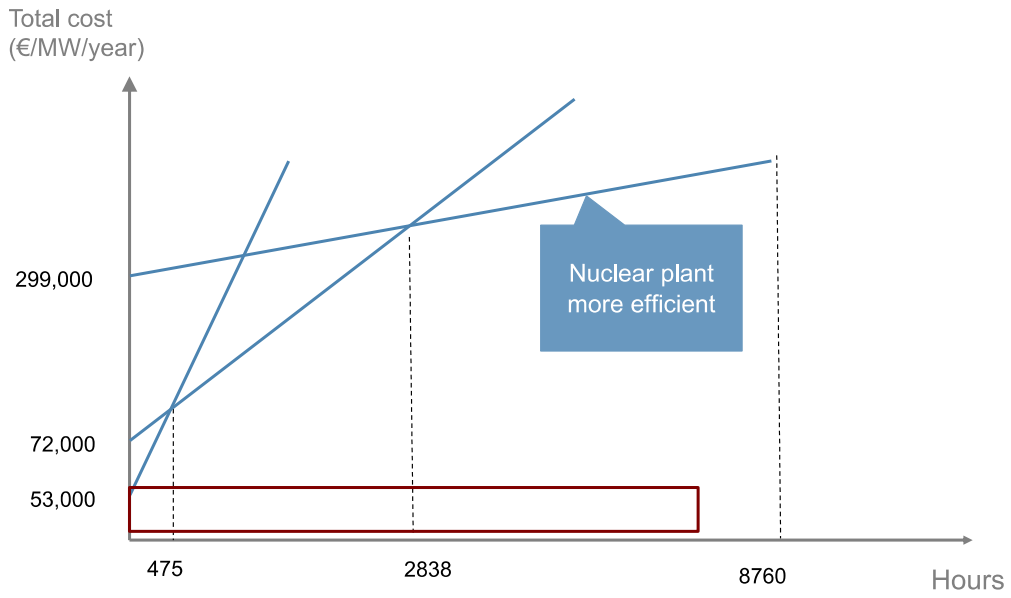

Figure 12: Optimal usage of a nuclear power plant. *Source: Léautier, 2019, figure 2.10, page 42.*

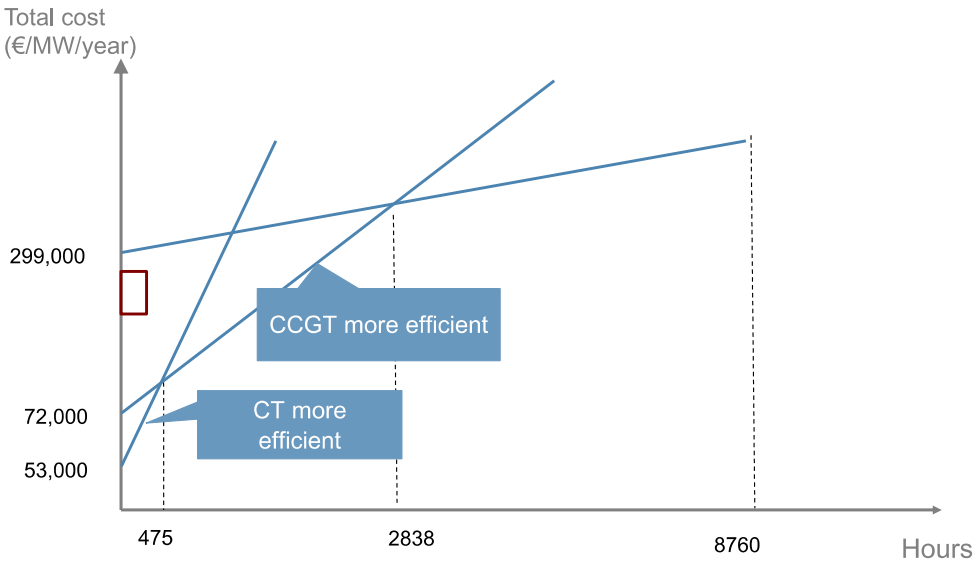

Figure 13: Optimal usage of an open cycle gas turbine (OCGT or CT). Source: Léautier, 2019, figure 2.10, page 42.

3.2.3 Resulting Supply or Dispatch Curve

Let us now turn to the generation supply curve (or dispatch curve), presented in Figure 14. The first flat portion corresponds to the hours when nuclear is the only technology producing. The price is thus equal to the variable production cost of nuclear, which determines demand. Then comes the first vertical portion of the supply curve: demand is equal to nuclear production, which is equal to nuclear capacity and the price rises to precisely balance demand and nuclear capacity. If nuclear was the only technology present, this would be the end of the story.

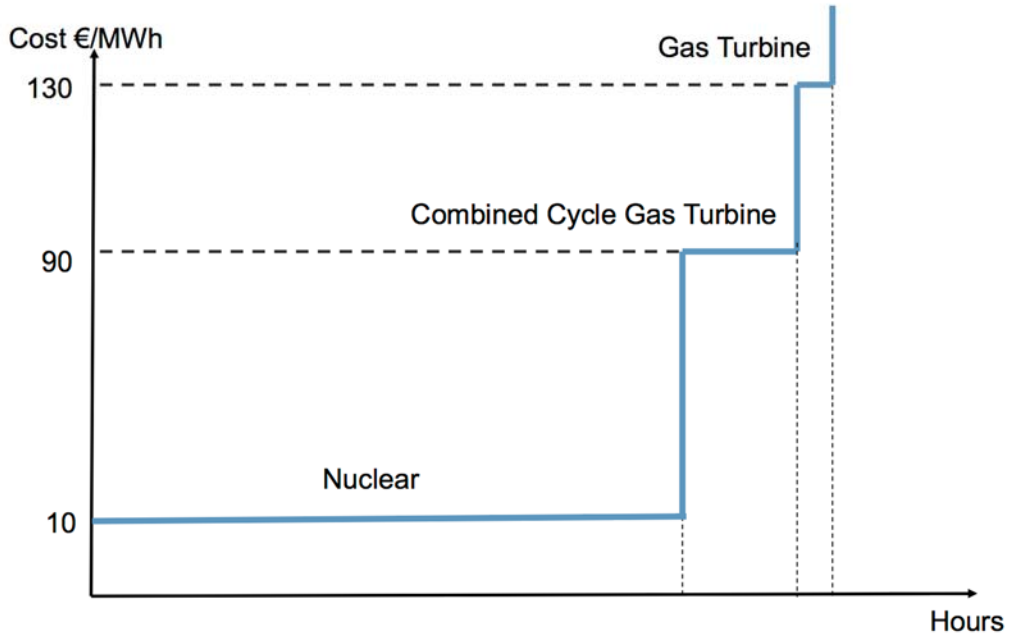

Figure 14: Supply curve for three technologies. Source: Léautier, 2019, figure 2.11, page 42.

However, since another technology is present, the story continues. The CCGT is turned on as soon as demand is high enough that the price exceeds its variable production cost, that is, when users are willing to pay more for a megawatt-hour than the cost of gas to produce it. Then, we travel the second flat portion of the supply curve: the price is equal to the variable production cost of CCGT, which determines demand.

Then comes the second vertical portion of the demand curve: price increases so that demand is equal to the cumulative nuclear plus CCGT capacity.

Finally, we travel the third L of the supply curve: a flat portion where the price is the variable production cost of OCGT, then a vertical portion where the price is such that demand is equal to the cumulative nuclear plus CCGT plus OCGT capacity.

3.2.4 Optimal Cumulative Capacity and Generation Mix

If demand responds to price, the logic presented in Section 2 also applies here to determine the optimal capacity and generation mix. The OCGT captures positive operating margin only when it produces at capacity, that is, when demand is equal to the cumulative capacity of all three technologies. In Figure 15, this is the last vertical segment of the supply curve, starting from point A_3 on the right of figure 15 (all technologies produce at capacity).

The cumulative capacity K_3 (on the right of the horizontal axis on Figure 15) installed in the long-term equilibrium is, therefore, such that the OCGT's average hourly operating margin is exactly equal to its hourly fixed cost. It is solely determined by (i) the cost of the marginal technology, in this case the OCGT, and (ii) the demand function. Thus, the long-term equilibrium cumulative installed capacity K_3 does not depend on the entire generation mix, but only on the marginal (peaking) technology.

Once the cumulative capacity has been determined, we need to define the long-term equilibrium generation mix. By construction, each technology produces if and only if its operating margin is non-negative. In Figure 15, the CCGT, for example, captures positive operating margin on the segments [A_2, B_2] (CCGT produces at capacity, OCGT not yet turned on), [B_2, A_3] (CCGT produces at capacity, OCGT produces partially), and the last vertical segment of the supply curve, starting from point A_3.

The free entry conditions for each technology are: average hourly operating margin, generated during the fraction of hours it produces at capacity, is exactly equal to hourly fixed cost. For example, the long-term equilibrium cumulative nuclear plus CCGT capacity K_2 is such that the CCGT operating margin (from point A_2 onwards) is equal to the hourly CCGT fixed cost. It does not depend on the cost of the nuclear capacity.

The same argument applies to the nuclear technology, that produces at every hour. The long-term equilibrium nuclear capacity K_1 is such that the average nuclear operating margin when nuclear produces at capacity is equal to the hourly nuclear fixed cost.

An essential feature of the long-term equilibrium is that the free entry condition applies to all technologies, that is, the three technologies precisely break-even.

Suppose the generation mix is at the long-term optimum and the system operator decides to impose a price cap, higher than 130 €/MWh (otherwise the OCGT would never be turned on), but low enough to be binding in some states of the world. The operating margin of the OCGT will be reduced and will be lower than the fixed hourly cost of OCGT. The operating margins of all technologies will also be reduced by the same amount and fall below the hourly fixed costs. This property is very important for market design: if the peaking technology is "missing money" at the long-term equilibrium, then all technologies are also "missing money" by the same amount.

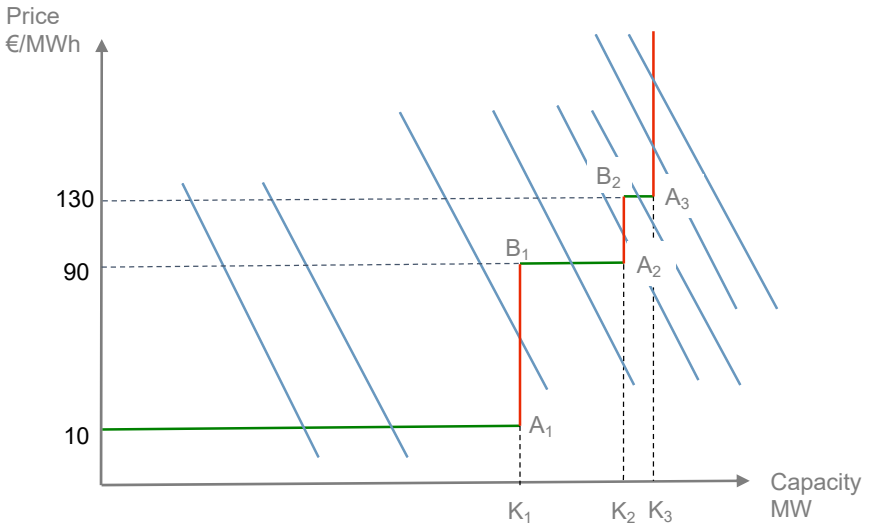

Figure 15: Supply and demand curves for three technologies.

Finally, the free entry conditions for inframarginal technologies can be interpreted as: the equilibrium mix is such that substituting one megawatt of one technology by one unit megawatt of the next technology generates no gain nor loss in total generating cost on the margin.

3.2.5 Optimal Cumulative Capacity and Generation Mix if no Customers Respond to Price

Power engineers, who historically assumed demand does not respond to prices, determined the optimal capacity and generation mix by combining the load duration curve (Figure 1) and the screening curves (Figure 11). This analysis is presented in Figure 16.

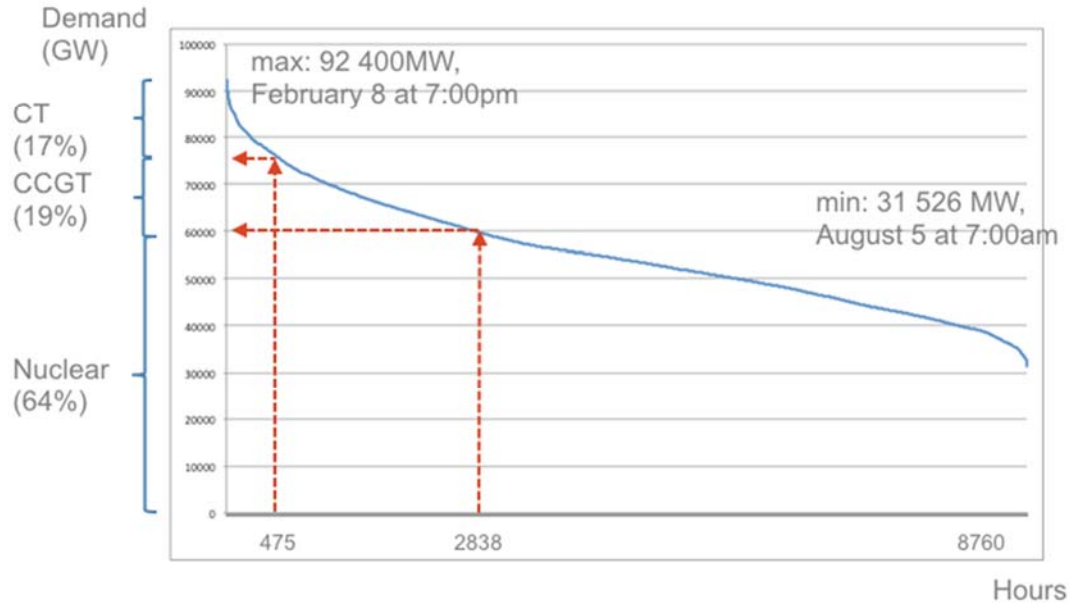

Figure 16: Determination of the optimal generation mix combining (i) the screening curves and (ii) the load duration curve. Source: Léautier, 2019, figure 2.12, page 43.

First, cumulative capacity is determined to meet peak demand, up to the generation adequacy criterion, for example, 116 per cent of expected peak demand.[19]

Second, analysis of the screening curves in this example shows that nuclear is the most efficient technology to serve a strip of demand lasting more than 2 838 hours. As discussed in Section 1, the load duration curve shows that demand exceeds 60 000 MW for 5 676 half-hours of the year, which is exactly 2 838 hours. An equivalent formulation is that the size of demand strips lasting more than 2 838 hours is 60 000 MW. Therefore, optimal nuclear capacity is 60 000 MW or 64 per cent of installed capacity.

Similarly, the load duration curve shows that demand exceeds 77 000 MW for 475 hours per year or equivalently the size of demand strips lasting more than 77 000 MW is 475 hours. Therefore, the optimal CCGT capacity is 77 000 – 60 000 = 17 000 MW or 19 per cent of peak demand.

Finally, OCGT constitutes the remaining generation capacity, 17 per cent of peak demand.

Finally, Figure 16 illustrates that nuclear capacity (for example) exceeds nuclear generation 5 676 non-consecutive half-hours per year. During these hours, nuclear produces at less than full capacity.[20] These hours correspond to the first flat portion of the supply curve (Figure 15).

3.2.6 Invariance of the Long-run Average Price

We have seen that the baseload technology runs all the hours. Thus, a baseload producer receives the (time-weighted) average price over the year. If this average is higher than the hourly long-run marginal cost, an additional unit is added into the market; hence the average price is reduced. This process is repeated until the average price is exactly equal to the hourly long-run marginal cost. Conversely, if the average price is lower than the hourly long-run marginal cost, baseload installed capacity is retired until the time-weighted average price is exactly equal to the hourly long-run marginal cost. This is the free entry condition. The average price is entirely determined by the hourly long-run marginal cost of the baseload technology.

This property holds exactly for the long run optimum. Industry equilibrium at any time differs from this long run optimum. Still, this property is directionally correct: if the time-weighted average price is today and is expected to remain for the next few years below the hourly long-Run marginal cost of the baseload technology, investors in baseload assets will tend to retire them.

This property is very important and produces counter-intuitive results. For example, increasing the share of price responsive customers, while increasing the net surplus, does not modify the time-weighted average price in the long term.

Similarly, since the mid-2000s, RES producers (for example, wind and solar) have been subsidised, hence have massively entered into electricity markets. Since their marginal operating cost is essentially zero (wind and sun are free) and their fixed cost is subsidised, one would expect the average price to be significantly reduced. And indeed, this is what we observe in Europe. However, this average price reduction is a short-term effect. Since the operating margin is lower than the fixed cost and expected to remain so for the foreseeable future, a fraction of the installed capacity is being retired, hence the average price will increase.[21] This process will stop precisely when the average price equals the hourly long-run marginal cost of the baseload technology, in this case nuclear. Thus, renewables entry has no impact on the average time-weighted wholesale price in the long term equilibrium, which is determined by the cost of the baseload technology. This property holds until baseload technology is pushed entirely out of the market by intermittent generation.

3.2.7 Long-run Marginal Costs

We have seen that long-run marginal costs are higher for a CCGT and an OCGT than for the baseload nuclear plant. We have also seen that the time-weighted average price is exactly equal to the long-run marginal cost of baseload technology for an optimal system in long-run equilibrium. Therefore, the time-weighted average price is strictly lower than the long-run marginal costs of a CCGT and an OCGT. If they were producing all the time, a CCGT and an OCGT would not cover their fixed cost at an average price equal to the long-run marginal cost of baseload generating capacity. How can they cover their cost by operating less than full time? The answer is that these technologies operate for fewer hours than the base-load technology but these hours have higher than average prices. The number of hours a plant is operating is not sufficient to determine its profitability. What matters is the number of hours a plant is producing at capacity, which are the only hours when it captures positive operating profit and the price during these hours.

3.2.8 Is the Generation Mix ever Optimal?

The analysis above appears simple enough. In reality, the generation mix is, of course, never optimal, for three main reasons, which are not specific to the power industry.

First, generation assets last more than 20 years. Investment is decided today, based on assumptions of future screening curves (that is, fixed and variable costs of generation technologies) and future load duration curves. These assumptions are almost always wrong: fuel prices change, taxes or subsidies are decided that impact costs, demand grows more or less than expected, etc.

Second, adjustment is not easy. When excess capacity has been installed, it is not immediately shut down, as has been observed in the NorthEast of the United States in the early 2000s and in Europe since 2010. Investors may keep assets running even if they do not fully recover their cost of capital, as long as they cover their variable costs. This decision is consistent with economic theory, which suggests that, since sunk costs are sunk, investors prefer to run an asset (as opposed to shutting it down) as long as its operating margin is positive (more precisely, as long as it generates a positive free cash flow). Thus, as long as an existing generating unit expects to cover its going-forward costs (fuel, O&M, including incremental capital costs), it is rational to continue operating rather than retiring it. When expected future revenues fall below expected future operating costs, it is rational to retire the generating plant.

In reality, assets are financed by debt and equity. When the free cash flow falls below the level required to service the debt (interest payment and principal repayment), the company files for bankruptcy and assets are restructured. During the restructuring process, a fraction of the profitable assets is may be retired.

Also, decisions could be less rational, for example, attributable to regulatory or political considerations.

Third, we have assumed throughout this chapter that industry equilibrium is optimal. In reality, this is not true. Investors do not necessarily coordinate to reach the optimal generation mix. They have a tendency to over-invest when prices are high and under-invest when prices are low. This is known as the “boom-bust” investment cycle.

3.3 Congestion on the Transmission Grid

So far, we have ignored the spatial dimension of power markets: we have considered only one location, where production and consumption both occur. In reality, power markets are spread geographically. Production often occurs far away from consumption centres. These different locations are connected to a high voltage transmission grid that transmits electric power over hundreds and sometimes thousands of miles. For example, a single transmission grid connects all of continental Europe, that is, all generation units and consumption centres in continental Europe are connected to the same grid and synchronised. In North America, three grids exist: the Western and Eastern grids and a separate grid for most of Texas (ERCOT). All generation units and consumption centres in the Eastern United States and Canada (roughly west of the Rocky Mountain States and Canadian Provinces) are connected to the same synchronised grid.[22] Similary in the West, with British Columbia and Alberta connected to the Western network.

This section discusses briefly how the peak-load pricing logic presented in Section 2 applies to an interconnected grid. Readers are referred to Chapters 3 and 6 in this Handbook, Léautier (2019) Chapter 6, Schweppe et al. (1988) and Hogan (1992) for more detailed discussions of transmission networks, congestion and locational pricing.

The main issue considered here is congestion on the transmission grid.[23] Power flows on transmission lines are limited. These limits arise for two reasons. First, there are thermal limits: if power flowing on a line is too high, the line will heat up and may break. Alternatively, the line will sag and may touch nearby trees, which would produce a short-circuit.

Second, there are operating limits. If a power plant or another line on the network fails, power flows are instantaneously rearranged, following the laws of physics. The operating limit on each line is such that, in the event of one (or more) failure on the system, the resulting flow on this line does not exceed the physical limit. This is called the (N-1) criterion or the single contingency rule: the system typically is operated to withstand the loss of one major component. Some system operators use an (N-2) criterion and operate their system to withstand the loss of two major components. For this reason, operating limits are often much lower than thermal limits.

Thus, a power market can be viewed as a series of "power islands" linked by bridges of limited capacity. When the traffic is low, it flows freely; the islands are all connected; hence a single market exists and the analysis presented in Section 2 applies.

When traffic is high, congestion sets in, the islands are separated and different markets exist. The analysis presented in Section 2 applies within each market: the local price is determined by the intersection of supply and demand (including import and export). When a technology is marginal, the local price is the variable cost of this technology. When a technology produces at capacity, the local price is the value of the marginal megawatt-hour consumed, up until the next technology starts producing.

The peak-load pricing logic applies to interconnected markets in another, more subtle, way. When there is no congestion between two markets, prices in each market are equal; hence the price for transmitting one megawatt-hour from one market to the other is zero, which is the marginal cost of transmission.[24]

When congestion is present between two markets, prices differ. To transmit one megawatt-hour from the low price to the high price market, a market participant injects one megawatt-hour in the low price market and withdraws it in the other. Thus, she receives the low price and pays the high price.

This leads to two observations. First, the price of transmission services is defined implicitly and is equal to the difference in local electricity prices. Second, the peak-load pricing logic presented in Section 2 applies to the pricing of transmission capacity: when the grid is not congested, the price of transmission service is equal to the variable cost of transmission, in this case, zero. When the grid is congested, the price of transmission service increases above the variable cost of transmission and is determined by the users' marginal valuation, which is the difference in locationalprices.

3.4 Inter-temporal Linkages and Operating Reserves

The previous story does not specify how the system adjusts in real-time to supply or demand shocks. We have simply said, 'in each hour demand is equal to supply, since limited storage is available today'. We have therefore implicitly assumed that production can adjust perfectly to changes in demand.

In practice, it is not that simple. Most production facilities have limited abilities to adjust. For example, the ramp-up rate determines the speed at which a power plant can increase its production and the ramp-down rate the speed at which it can decrease its production. Ramp-up and ramp-down capabilities vary by technologies, some being more flexible than others. In addition, starting up and/or shutting down a power plant is costly.

Furthermore, supply and demand are subject to sudden random shocks. For example, a power plant may "trip" and suddenly stop production, the wind may exceed the acceptable speed and wind turbines may suddenly shut down, a large user may unexpectedly stop his production process, hence his electricity consumption. This would not be a problem if production facilities had unlimited ability to adjust, but adjustment is limited.

How do power systems cope with the presence of random shocks and limited ramp-up and down flexibility? First, system operators and power producers do not make production decisions based on a single hour, rather based on a stream of hours: production at an hour $(t+1)$ is partially determined by production at hour t. For most power plants, the dispatch decision can be decomposed in two related decisions: (i) decide today whether to turn the plant on tomorrow, a decision known as unit commitment, then (ii) decide how much to produce for every hour tomorrow, taking into account the ramp-up and down rates, as well as the minimum production level required by the machine. This two-stage unit commitment problem is significantly harder to resolve analytically. Fortunately, a branch of applied mathematics, called Operations Research, is dedicated to solving these kinds of problems. The economic intuition is basically unchanged, although the analysis is much richer.

Second, system operators create operating reserves as a "cushion" to respond to these kinds of shocks. Operating reserves are extremely important operationally, as the lights would go out otherwise. From an economic perspective, they can be treated as additional demand, that is, the system operator demands operating reserves in the same way a user demands electric power. Thus, operating reserves do not modify the standard peak-load pricing story.

3.5 Three Time-horizons of "Security of Supply"

Policymakers and practitioners often mention "security of supply" when discussing the electricity industry, usually in sentences such as 'the Government will guarantee security of supply'. This term is confusing, since it mixes three different time-horizons, hence three different notions. It is essential to disentangle them.

3.5.1 Energy Security of Supply

Energy security of supply is the ability of a region (for example, a country, a group of countries or a state in a federal country) to secure its long-term supply of primary energy. It can be crudely measured in units of energy: terawatt-hours (TWh), tons of oil equivalent (TOEs), or gigajoules (GJ).

The simplest form of security of supply is to own the primary energy required to fuel the economy for the foreseeable future. A more sophisticated approach is to have long-term contracts or agreements with "friendly" foreign governments.

For example, when the US President Franklin Roosevelt met King Abdulaziz of Saudi Arabia on board the US Navy cruiser Quincy on 14 February 1945, he entered into a security of supply agreement: the US would protect Saudi Arabia and, in return, Saudi Arabia would continue to export its oil to the US. This agreement was particularly important during the 1970s when Saudi Arabia was a force stabilising OPEC supply and prices (Yergin, 1990).

In 1973, faced with the oil crisis and the risk of oil supply constraints and price spikes, the French government launched the nuclear electricity program to guarantee the security of its power supply by replacing electricity produced with petroleum with nuclear electricity and displacing oil used by consumers with electricity (electrification). The primary fuel, uranium, was procured from current or former French colonies (New Caledonia in the Pacific Ocean and Niger in Africa).

Can competitive firms guarantee energy security of supply? Theoretically yes. In practice, however, probably no. While private firms sign long-term commercial contracts, these contracts are often linked to broader alliances between countries. Energy security of supply falls squarely within the government's purview.

3.5.2 Generation Adequacy

As we have seen above, when electricity demand does not respond to short-run price variations, which was the case for most of the twentieth century, generating capacity is determined to meet demand for the next year, the next five years or the next ten years to achieve in expectation an agreed-upon reliability criterion. The capacity that satisfies the generation adequacy standard is measured in units of electric power (usually GW). It is routinely estimated and reported by (transmission) system operators.

Generation adequacy is often confused with security of supply in the public discourse. This is misleading. The former is ensuring that, on average, rolling blackouts do not exceed an agreed-upon level, for example, three hours per year. If the generation adequacy standard is not met, rolling blackouts may reach ten hours during a very cold winter in Europe (or a very warm summer in the United States) or brownouts may occur. This is, of course, unpleasant. No one likes to be deprived of electric heating precisely when the temperature drops. But it is much less dramatic than having to ration all users for the entire winter because the gas reserves in the country are insufficient.

3.5.3 System Reliability

System reliability is the ability for the system to react in real-time to unforeseen circumstances, for example, a sudden loss of generation, a sudden demand increase, or any event that materially and suddenly affects power flows.

Remember that power injected in the grid must equal power taken out of the grid (either for consumption or dissipated through losses or stored) at all times. Excess demand causes a frequency drop and reciprocally excess supply causes a frequency increase. If this deviation exceeds a certain very tight level, parts of the power system automatically shutdown. This would yield an uncontrolled blackout or system collapse. In August 2019, Great Britain experienced an

uncontrolled blackout due to two power plants shutting-off simultaneously; hence the frequency dropped below 48.8 Hertz. This triggered National Grid's (the system operator in Great Britain) automated system to cut off supply to around 5 per cent of demand. One million customers, including train stations and hospitals, lost power for a few hours.

This must be distinguished from an organised rolling blackout reflecting an ex-ante plan to respond to inadequate generating capacity to balance supply and demand. In the latter, the underlying economics is that it is more efficient (ex ante) to curtail customers for a few hours than to build an additional power plant. An uncontrolled blackout, on the other hand, is uneconomic. Buyers and sellers are prevented from executing economic transactions. One attribute of uncontrolled system blackouts is that it takes time (and effort) to recover. A power grid that has collapsed is not simply switched back on with the flip of a switch. The 2003 blackout in the Northeastern US is a good example of a system collapse and recovery, though the cause of the system collapse was a transmission outage rather than a generation shortage.[25]

To guarantee the reliability of the power grid, system operators have put in place protocols and practices. As previously discussed, SOs procure operating reserves to guarantee system reliability. Another example: the maximum power allowed to flow on transmission lines is lower than the physical capacity of the line so that the system is able to operate even if one large component (for example, a large power plant or another transmission line --- the largest contingency) suddenly fails – the N-1 criterion. What matters for reliability is the ramp rate (up or down) at which generators can adjust their output. It is expressed in MW per minute.

The economics of reliability constitutes an area for fertile future joint work by economists and engineers. First, most reliability protocols and practices are very static, which were designed before the condition of system components could be measured and analysed in real-time. As sensors and remote control devices are installed on the power grid, these reliability protocols and practices will become dynamic, hence more efficient. Second, in the (near) future, demand response and batteries will become a non-negligible reality in power markets and will contribute to system reliability. Robust rules to include them will have to be designed.

3.5.4 Additional Thoughts

3.5.4.1 The Link between the Three Time-horizons

The three time-horizons are not as separate as the previous discussion may suggest. They feed continuously into one another. For example, generation adequacy can be estimated the day before or even two hours before actual dispatch, hence resembles system reliability. However, they are different notions, expressed in different units. Poor security of supply leads to long episodes of shortages, usually administered through rolling blackouts. Inadequate generation adequacy leads to more rolling blackouts than is efficient. Inadequate reliability leads to unmanaged system blackouts.

3.5.4.2 Comparison with Other Industries

Observers sometimes argue that reliability is unique to the power industry: 'if there is a shortage of planes in Chicago, it does not affect travellers on the East Coast'. The distinction between capacity adequacy and system reliability helps clarify this point.

Adequacy is having sufficient infrastructure to accommodate the next anticipated spike in demand, for example having sufficient planes in Chicago next summer. If planes are short next summer in Chicago, prices will go up, some travellers will not be able to fly on their preferred date and some may find themselves on the wrong end of overbooking. This is indeed unlikely to affect passengers on the East Coast.

On the other hand, as all US air travellers know, if thunderstorms hit Chicago in the summer, as they often do, flights in and out of Chicago will be delayed, possibly cancelled, which will most definitely affect travellers on the East Coast (and on the West Coast also). The consequences of a local reliability issue are less severe for the air travel system than for the power grid. In particular, total blackouts are extremely infrequent. The difference, however, is a matter of degree, not principle.

4 Concluding Remarks

This chapter has shown that the structure of wholesale power prices is relatively straightforward: prices are close to the variable cost of production during off-peak hours, when capacity exceeds demand, and are set by the valuation of the marginal consumer when demand is exactly equal to capacity. On-peak prices can be 20 to 50 times higher than off-peak ones. We have developed this basic result and other important results that it implies by introducing increasingly realistic models.

There are several relevant economic considerations that we do not address in this chapter. We assume implicitly that there is no market power in wholesale markets: suppliers are price takers. Market power considerations are introduced in Chapter 3 and Chapters 5-10 discuss, among other things, how market power has been addressed by these different market designs. Nor do we discuss the variations in the mechanisms that individual wholesale market designs have employed to implement these basic principles. Chapters 5-10 of this Handbook do so.

The discussion in this chapter also embodies the assumption that generating plants are dispatchable by the system operator. That is, the system operator can effectively control the economic dispatch of generators that have been scheduled to supply through the prevailing wholesale market mechanisms. However, in response to policies to decarbonise the electric power sector, traditional dispatchable generation is being replaced by wind and solar (primarily PV) technologies. The supply of electricity from these technologies is "intermittent," depending on uncontrollable variations in wind and sun. Furthermore, distributed generation (DG) on the customer side of the meter, primarily rooftop PV, rather than on the transmission grid, is also expanding quickly in many countries. The large-scale diffusion of these intermittent technologies, combined with government policies that are replacing market mechanisms with administrative obligations, is creating new challenges for the design of efficient wholesale markets (Joskow 2019). How these challenges can be addressed is discussed in Chapters 14 and 15 in Part II of this Handbook. One technological response to intermittency is investments in energy storage technologies which can essentially shift electricity supplied when it is cheap to periods when it is dear, subject to a variety of storage capacity, cycle-time, charge and discharge constraints. Storage technologies are sometimes consumers of energy and sometimes producers of energy. Efficient integration of storage opportunities into wholesale markets is another challenge. Finally, new metering and remote sensing and control technologies are creating new

opportunities to create a much more active demand side that allows consumers to respond more quickly and effectively to price changes and to express their willingness to pay for reliability. These opportunities are discussed in Chapter 12.

Whether it is the relatively simple models considered in this Chapter or enhancements to accommodate the changes noted above, there is one fundamental pricing principle that continues to prevail. This is the role of prices that vary widely with variations in supply and demand. Producers capture their highest profits when consumers desperately need electric power to heat (or freshen) their houses. This is likely to be even more important for producing market revenues sufficient to cover the total costs of new generating technologies, as the short-run marginal costs of wind and solar are essentially zero. If the market price is zero, no net revenues are produced to cover generators' capital costs. Thus, scarcity pricing – incidents of very high prices necessary to clear the market – must play a more important role in the future to satisfy generators balanced budget constraints. Economists argue that this outcome is perfectly acceptable; in fact, it is optimal. Consumers and their elected representatives have a different opinion. They argue that profiteering from consumers' need is amoral, hence unacceptable.

Resolving this tension is essential to the future of the power industry. Consumers (some at least) and policymakers are looking forward to the decentralisation of the power industry: consumers equipped with green and decentralised generation and storage (a Tesla car in their garage) will be active participants in the power markets. How do we coordinate the decisions of millions of economic agents? Prices reflecting the value of power at every instant and every location seems the most natural approach. This requires policymakers and consumers to reconcile themselves with possibly extremely high prices at some instants in some locations. Otherwise, the decentralization of the electricity system will prove an elusive goal.

References

Boiteux, Marcel (1960) (English Translation), 'Peak Load Pricing', *Journal of Business*, April, **XXXIII** (2), 157-179. Originally published in French in 1949.

Boiteux, Marcel (1951), 'La Tarification au Coût Marginal et les Demandes Aléatoires !, *Cahiers Seminaire d'Econométrie*, **1**, 56-69.

Boiteux, Marcel (1956), 'Le Choix des Équipments de Production d'Énergie Électric', *Review Francais de Recherch Opérationelle*, **1** (1), 45-60.

Borenstein, Severin (2005), 'The Long Run Efficiency of Real-Time Electricity Prices', *Energy Journal*, **26** (3), 93-116.

Cramton, Peter and Steven Stoft (2005), 'A Capacity Market that Makes Sense', *The Electricity Journal,***18**, 43-54.

Dreze, Jacques (1964), 'Some Post-war Contributions of French Economists to Theory and Public Policy: With Special Emphasis on Problems of Resource Allocation', *American Economic Review*, **54** (4), part 2, Supplement, 2-64.

Hogan, William (1992), 'Contract Networks for Electric Power Transmission', *Journal of Regulatory Economics*, **4** (3), 211-242

Joskow, Paul L. (1996), 'Introducing Competition into Regulated Network Industries: From Hierarchies to Markets in Electricity', *Industrial and Corporate Change*, **5** (2), pp 341-382.

Joskow, Paul L. (2007), 'Competitive Electricity Markets and Investment in New Generating Capacity', in Dieter Helm (ed.), *The New Energy Paradigm*, Oxford: Oxford University Press.

Joskow, Paul L. (2008), 'Capacity Payments in Imperfectly Competitive Electricity Markets', *Utilities Policy*, **16**, 159-170.

Joskow, Paul L. (2019), 'Challenges for Wholesale Electricity Markets with Intermittent Renewable Generation at Scale: The U.S. Experience', *Oxford Review of Economic Policy,* **35** (2), 291-331.

Joskow, Paul L. and Jean Tirole (2007), 'Reliability in Competitive Electricity Markets', *Rand Journal of Economics*, **68**, 159-170.

Keppler, Jan Horst (2017), 'Rationales for Capacity Remuneration Mechanisms: Security of Supply, Externalities and Asymmetric Investment Incentives', *Energy Policy* **162**, 562-570.

Léautier, Thomas Olivier (2016), 'The Visible Hand: Insuring Optimal Investment in Electric Generation', *Energy Journal*, **37** (2), 89-109.

Léautier, Thomas Olivier (2019), *Imperfect Markets and Imperfect Regulation*, Cambridge, MA and London, England: MIT Press.

Newbery, David, Michael G. Pollitt, Robert A. Ritz and Wadim Strielkowski (2018), 'Market Design for a High-Renewables European Electricity System', *Renewable and Sustainable Energy Reviews*, **91**, 695-707.

Schroder, Thomas and William Kuckhsinrichs (2015), 'Value of Lost Load: An Efficient Economic Indicator for Power Supply Security? A Literature Review', *Frontiers in Energy Research*, **3**, 1-12.

Schweppe, Fred, Michael Caramanis, Richard Tabors and Roger Bohn (1988), *Spot Pricing of Electricity*, Norwell, MA: Kluwer Academic Press.

Turvey, Ralph (1968), *Optimal Pricing and Investment in Electricity Supply*, London: George Allen and Unwin.

United States Energy Information Administration (US EIA) (2020), 'Levelized Cost and Levelized Avoided cost in the *Annual Energy Outlook 2019*', accessed on September 9, 2020 at www.eia.gov/outlooks/aeo/pdf/electricity_generation.pdf.

Wolak, Frank (2018), 'Efficient Pricing: The Key to Unlocking Radical Innovation in the Electricity Sector", manuscript July 22.

Notes

[1] This chapter draws heavily on Chapter 2 of Leautier (2019). Mathematical derivations of the primary results discussed here can be found there as well.
[2] Extensions of these models to integrate intermittent or non-controllable generating capacity and electricity storage are discussed in Chapter 1, Chapter 14, and Chapter 15. See also Léautier (2019) Chapter 8, Joskow (2019), Newbery et al. (2018) and the references they cite.
[3] An additional difference between the wholesale spot price for power and the retail price for hotel rooms or plane seats is that hotels and airlines also are able to price discriminate among users: two passengers seated next to one another may have paid vastly different prices for their seat. This type of price discrimination does not exist in wholesale spot markets where "the law of one price" holds at any point in time. A more precise formulation of the above statements would be "the minimum price for a room in the winter is close to the cost of cleanup".
[4] We use euro as the currency in most of the chapter but we sometimes use $/MWh and £/MWh when referencing data or studies that use US$ or British Pounds. Given current exchange rates, all three units are roughly equivalent.
[5] The United States has a large number of system operators, sometimes vertically integrated utilities and sometimes Independent System Operators (ISO), spread over a country with four time zones, variations in customer composition, and variations in weather. The time of peak demand for each system operator varies. Thus, the peak demands of the system operators are "noncoincdent" and adding them all up yields the noncoincident peak demand. This is how the aggregate data for the United States is reported.
[6] The load duration curve is still representative, as it has not changed significantly in the last ten years.
[7] Multiple generating technologies are introduced into the model below.
[8] Numbers are expressed according to the following convention: a thin space separates the tens of thousands and a dot separates the integer from the decimal parts.

[9] Fixed and variable costs of producing electricity vary over time, since technologies, interest rates and commodity price evolve. The costs used here are purely illustrative. For example, the variable cost in the US is currently lower than 50 $/MWh because natural gas and coal are less expensive than in Europe and Asia. The US Energy Information Administration reports a levelised capital cost of about $US (2018) 75 000 per MW for a conventional combustion turbine. See US EIA (2019) Table 1b. Levelised fixed O&M costs add another $US (2018) 18 000 per MW.
[10] The efficiency of real-time prices that reflect the marginal cost of supplying electricity, including scarcity prices (more below), is well established in the literature. For example, Borenstein (2005) and Wolak (2018).
[11] Many residential and small commercial consumers in some countries still are faced with a bundled price that includes both wholesale energy costs and delivery (transmission and distribution) charges. We will ignore this complication in this discussion.
[12] In the US, industrial customers typically avoid buying directly in the wholesale market for a variety of legal, regulatory and practical reasons. They rely on wholesale market intermediaries which compete to be their electricity supplier.
[13] Actual curtailment plans are more complex than this. Typically, the SO or a government agency will first issue a public call for conservation by consumers in light of a pending emergency. Next, the SO may allow operating reserves to decline below their target level. Then, the SO may reduce the voltage on the system ("brown-out"). Finally, the SO will turn to rolling blackouts or curtailments to reduce demand on the system.
[14] Regulators may require system operators to apply price caps below VoLL to wholesale market realisations in order to mitigate real or imagined market power problems in wholesale markets. However, when *competitive* market clearing prices exceed the price cap non-price rationing of supply is necessary to balance supply and demand. Binding price caps in competitive wholesale markets can have significant effects on short-run and long-run (investment) market efficiency. The rules for price caps (or offer caps) for system operations in the US are specified in Federal Energy Regulatory Commission (FERC) Order 831 – (2016), available at https://www.ferc.gov/legal/maj-ord-reg.asp?new=sc3.
[15] The short run marginal cost of generating electricity from a wind and solar plant is roughly zero. If wind and solar facilities with specific generating capacities were dispatchable, an economic dispatch would lead the system operator to call on this capacity whenever the aggregate specified wind and solar capacity is greater than or equal to demand. However, wind and solar generation is intermittent and it's availability depends on variations in the wind and sun. Thus, its effective generating capacity varies based on the intensity of the wind and the sun at any particular time.
[16] New Zealand is largely hydro-based. The main supply risk is represented by a dry year during which rolling blackouts would occurs for a few winter days.
[17] In practice, ERCOT and other energy-only markets also design a schedule of administrative procedures and out-of-market actions they employ before initiating rolling blackouts.
[18] California is also an exception. Since 2001 the California Public Utilities Commission has placed resource adequacy requirements on load serving entities, until recently primarily distribution utilities, whether covered by the California ISO or not, rather than on a centralized capacity market administered by the ISO. This is a rather bazaar system that leads to all kinds of problems especially as retail competition is introduced. California has been considering moving to a centralized capacity market mechanism. More information is available at https://www.cpuc.ca.gov/RA/.
[19] To simplify the example, we choose installed capacity equal to peak demand and round up the numbers.
[20] In real power markets, nuclear plants schedule their maintenance during these hours, so that the active cumulative capacity follows the load curve. In addition, they may export to neighbouring markets.
[21] The dynamics here are complex, as discussed later.
[22] However, the Eastern and Western Interconnections have several balancing authorities with operating responsibilities for portions of each of these grids. ERCOT has one balancing authority. See https://www.eia.gov/todayinenergy/detail.php?id=27152 .
[23] To simplify this introductory discussion, we ignore transmission losses, which does not alter the basic economic analysis.
[24] Since transmission losses are ignored.
[25] https://en.wikipedia.org/wiki/Northeast_blackout_of_2003