Chapter 5:  
Optimal capacity mix and scarcity pricing

# Learning objectives

This chapter introduces a simple pen-and-paper technique that allows us to derive the optimal mix of thermal power plants (such as nuclear, coal-fired, and natural gas-fired plants) in the long run. It is based on screening curves (introduced in The cost of electricity) and so-called load duration curves. At the end of the chapter, we will be able to:

* derive the cost-minimal capacity mix,
* derive the cost-minimal generation (energy) mix,
* assess the impact of renewable energy on the mix,
* assess the impact of costs shocks, including carbon pricing, on the mix,
* understand peak-load pricing theory, and
* understand the “missing money problem”

# The load duration curve

The last chapter The price and value of electricity introduced short-term electricity market models that helped us understand how power plants are dispatched and how electricity prices are determined. The generation capacity, however, was simply there. In this chapter we look at how the decision to install that capacity can be taken. We introduce a simple technique that allows us to derive the cost-optimal generation capacity mix in an electrical system.

**Two pieces**. To derive the cost-minimal thermal generation mix, we need to ingredients: first, screening curves, which represent the cost of all available thermal generation technologies; these were introduced in the chapter on The cost of electricity. The second ingredient is the load duration curve (LDC), which describes the pattern of electricity consumption over time.

**Load duration curve**. We start with the “load curve” that displays electricity consumption of every single consecutive hour in Germany (**Figure 1**) during a year. Typically, electricity consumption is higher during the day than at night and higher during the week than on weekends or holidays, reflecting human and economic activity and needs. In cool climates, electricity demand tends to be higher during the winter, reflecting the demand for lighting and heating. These electrical systems are called “winter peaking systems”. In hotter climates, summer peaking systems prevail, where air conditioning causes electricity demand to peak on summer afternoons or evenings. The highest and lowest electricity demand during the course of one year can easily vary by a factor of two to three.

|  |
| --- |
| **Figure 1. German electricity consumption during 2010, displayed in consecutive order**  **Key point: Electricity consumption varies, with characteristic diurnal, weekly and seasonal fluctuations as well as random fluctuations.** |
|  |
| **Source:** OEE based on load data |

**The load duration curve**. The next step is to order electricity consumption by size. **Figure 2** shows the exact same data as **Figure 1**, just that the hours are ordered by decreasing load. This is the so-called “load duration curve”. One way to read it is to pick a point on the x-axis (say, 3000 hours) and determine the corresponding load level (about 62 GW). This means that in 3000 hours of the year, electricity consumption exceeded 62 GW. Another way to read the chart is to pick a point on the y-axis (say, 45 GW) and determine the corresponding number of hours (about 7500 hours). This means that in 7500 hours of the year the load did not drop below 45 GW. The extreme left point on the graph indicates that the peak load during the year was 80 GW and the extreme right point shows that load never fell below 35 GW.

|  |
| --- |
| **Figure 2. German electricity consumption during 2010, ordered by size**  **Key point: There are fewer hours during a year when the load is very high.** |
|  |
| **Source:** OEE based on load data |

# The optimal thermal generation mix

This section uses screening curves introduced in the chapter The Cost of electricity to derive the least-cost generation capacity mix. Note that our purpose is to introduce the optimization technique. The underlying screening curves used here are not necessarily representative of a real-world case.

**Optimal mix.** The least-cost generation capacity mix can be derived graphically by placing the relevant screening curves above the load duration curve in the same figure.

First, start at the very right of the screening curves panel in **Figure 3**. The screening curves tell us that nuclear power would be the cheapest way to generate electricity when a plant is required to run 7000 FLH or more, the point of intersection of screening curves of coal-fired and nuclear plants. Next, to determine the quantum of electricity demanded more 7000 hours in a year, draw a vertical line from the point of intersection onto the load duration curve. The vertical line intersects the load duration curve at a load of about 40 GW. This is the amount of nuclear capacity that should be built.

Proceeding in a similar manner we see that the next intersection to the left in the screening curves panel (between coal and natural gas) is at 2000 FLH, corresponding to 65 GW. The difference between 65 GW and 20 GW is the optimal capacity of coal-fired plants: 25 GW. Natural gas plants are the cheapest options for less FLH. The LDC indicates that this corresponds to only about 15 GW of capacity. The total of all the different technologies is about 80 GW (40 GW of nuclear plants + 25 GW coal plants + 15 GW of natural gas plants), which is just sufficient to meet the peak load.

|  |
| --- |
| **Figure 3. The cost-minimal capacity mix, as derived from screening curves and a load duration curve.**  **Key point: It is cost-optimal to have a mix of technologies.** |
|  |
| **Source:** OEE based on load data |

**A *mix* is optimal.** Let us briefly reflect on this result. Starting from scratch, knowing only the cost of different technologies and the load duration curve, we derived that it is cost-optimal to deploy *a range* of very different technologies to produce electricity in spite of one technology clearly being cheaper than others. In almost any other industry using the cheapest production process would be cost-optimal way to meet demand and, if you could optimize from scratch, you would employ only this one efficient technology. In these industries, presence of different production technologies usually reflects hedging in face of uncertainty or legacy production techniques. This is not what drives technology differentiation in case of electricity generation. In case of electricity, it is the combination of non-storability and varying load (that must be met at each moment in time) that results in a whole range of different technologies being simultaneously cost-efficiency.

**Optimal generation mix.** What is the value of electricity generation corresponding to the optimal capacity mix derived above? The area under the load duration curve gives total electricity consumption (and, hence, generation) during the year. The optimal generation corresponding to each technology is a subset of this area. For example the striped area in **Figure 4** represents the optimal annual generation of electricity from nuclear power and the dotted area shows the generation from natural gas-fired plants.

|  |
| --- |
| **Figure 4. Generation from nuclear plants (striped) and natural gas-fired (dotted)**  **Key point: Areas represent optimal level of generation for each technology.** |
|  |
| **Source:** OEE based on load data |

# The impact of cost changes

How does the capacity mix alter when the cost of different technologies or the pattern of load changes? Let us first consider the impact of change in costs using two examples: an increase in the investment cost of nuclear power plants (that increases the fixed cost of the power plants) and introduction of carbon pricing (that increases the variable cost of fossil fuel based power plants).

**Increase in fixed cost.** In recent years, many nuclear power plant projects have witnessed massive cost overruns. An infamous example is Finland’s Olkiluoto 3 power plant, the first of the new European Pressurized Reactor type, which is now expected to cost closer to EUR 8.5bn instead of the initial estimated cost of about EUR 3bn. How does a revision of cost play out in the optimal capacity mix? **Figure 5**illustrates this case. An increase in fixed costs results in an upward shift of the screening curve of nuclear power, while all other curves remain unchanged. Nuclear power is more expensive than coal even at 8760 full load hours. As a consequence, coal replaces nuclear in the long-term optimal capacity mix.

|  |
| --- |
| **Figure 5. Optimal capacity with increased cost of nuclear power**  **Key point: Nuclear is no longer competitive after the increase in cost.** |
|  |
| **Source:** OEE based on load data |

**Increase in variable cost.** In the past decades several jurisdictions have introduced prices on the emissions of CO2 from power generation, including several U.S. states, China and European Union. **Figure 6** shows the impact of introducing such a price on carbon. The cost curves of the fossil fuel based technologies rotate counter-clockwise. Further, the cost curve of coal rotates more than the natural gas curve because coal is more carbon-intensive. As a consequence, coal loses market share vis-à-vis both natural gas and nuclear. An even higher carbon price would lead to coal be not being competitive at all.

|  |  |  |  |
| --- | --- | --- | --- |
| |  | | --- | | **Figure 6. Optimal capacity with introduction of a carbon price**  **Key point: The share of coal in the optimal capacity mix declines.** | |  | | **Source:** OEE based on load data | |

# Impact of renewables on the optimal mix

In determining the optimal capacity mix, using screening curves of each technology, we only considered thermal generation technologies. This is because the screening curve technique does not serve well in case of variable generation of wind and solar power. Calculating the optimal capacity mix taking into account such renewable technologies requires us to use more complicated computer models. But the method described above can be used to understand the *impact* of introducing wind and solar energy on the cost-optimal thermal plant mix.

**Residual load duration curve.** To do this we use the concept of “residual load”, which we define as the electricity demand net of wind and solar generation. By sorting hours by residual load, instead of the actual load, we can derive the residual load duration curve (RLDC).

**Figure 7** displays RLDCs at different levels of penetration of wind and solar energy. While higher penetration rate only marginally reduces the y-intercept of the residual load duration curve, the curve pivots clockwise making it steeper than the load duration curve. Note that at a penetration rate of about 20 percent, the RLDC actually intersects the x-axis. This indicates that during some hours in the year, wind and solar plants supply sufficient electricity to serve the entire load.

|  |
| --- |
| **Figure 7. RLDC for different levels of wind and solar penetration**  **Key point: With increase wind and/or solar deployment, the RLDC become steeper.** |
|  |
| **Source:** Own work based on data taken from [Hirth et al. (2015)](http://dx.doi.org/10.1016/j.renene.2014.08.065) |

**Optimal thermal capacity mix.** Let’s get back to our optimal thermal capacity mix analysis. The expansion of renewable energy does change the cost curves. But the load duration curve is now replaced by the residual load curve. Further as the amount of wind and solar power in the system increases, the RLDC becomes steeper. This is shown in **Figure 8**.

|  |  |  |  |
| --- | --- | --- | --- |
| |  | | --- | | **Figure 8. Optimal capacity of nuclear power with increasing wind penetration**  **Key point: With growing renewable energy, less and less base load will be needed.** | |  | | **Source:** Own work based on data taken from [Hirth et al. (2015)](http://dx.doi.org/10.1016/j.renene.2014.08.065) | |

Note that the overall amount of thermal capacity required does not change much indicating a small “capacity credit” of wind and solar power. But the cost-optimal composition of capacity changes dramatically: a lower quantum of base load technologies (coal, nuclear) is needed, while the amount of mid and peak load technologies (natural gas) increases.

**Optimal thermal generation mix.** As explained above, by summing the areas under the RLDC one can determine the cost-minimal annual electricity generation mix. **Figure 9** shows the electricity generation by plants that operate as base load plants (8000 FLH or more), mid load plants (4000 to 8000 FLH), peakers (1000 to 4000 FLH) and super peakers (below 1000 FLH). In a traditional power system without wind and solar energy, it is cost-optimal to produce about three quarters of all electricity from plants that run almost all the time. When wind and solar reaches 40% penetration, it is no longer optimal to operate any power plant all the time. Electricity that is not generated by wind turbines and solar cells is more optimally produced in plants that operate only part-time (mid and peak load plants). **Figure 10** illustrates this point further. As the market share of renewable energy increases, less and less electricity is supplied by power stations that run base load. This transformation can either imply that power plants that were built for base load operation reduce their running hours or they be replaced by plants that are constructed for mid load operation.

|  |  |
| --- | --- |
| **Figure 9. Share of electricity generated from base load, mid load, and peak load plants without renewables and at 40% penetration.**  **Key point: As penetration of renewables increases, less electricity should be generated using base load plants and more using mid and peak load plants.** | **Figure 10. Share of electricity generated from base load plants (defined as 8000+ FLH) for different rates of wind plus solar penetration.**  **Key point: Base load disappears as renewable energy expands.** |
| |  | | --- | | **Source:** Own work based on data taken from [Hirth et al. (2015)](http://dx.doi.org/10.1016/j.renene.2014.08.065) | | |

# Scarcity or peak-load pricing model

**Recap: short-run model of electricity prices.** The merit order model introduced in the previous chapter was a short-run model i.e. a model that takes the existing capital stock (of power plants) as a given. In that model electricity price always equals the variable cost of the “marginal” power plant. A power plant earns short run profits if it is infra marginal i.e. if its variable costs are lower than the market price. But are the short run profits sufficient to cover fixed costs? Further, can the “marginal” power plant in the merit order never make a profit, as the price never rises above its variable cost? To answer these questions we need to employ a different, long-term perspective.

The starting point of long run price determination in electricity markets is the peak-load pricing or scarcity pricing theory.

**Scarcity pricing with inelastic demand**. In the simplest case, let’s consider a system with perfectly price-inelastic demand i.e. demand for electricity is the same irrespective of the price. **P**eak-load pricing theory shows that in this case in all hours of the year the market price for electricity will be equal to variable cost of the marginal plant *except in the hour with highest demand.* In that hour the price rises so high that the *entire* annualizedcapital cost can be financed during this one hour. For example, if annualized fixed cost of the marginal power plant is EUR 50,000 per MW (or EUR 50 per kW) the electricity price rises to recover this cost plus the variable cost of the plant. This could easily be several thousand euros per MWh. This is called “peak- load pricing” or “scarcity pricing”, because overall capacity is scarce in this, and only this, hour.

**Scarcity pricing with price-elastic demand**. What happens when the demand for electricity decreases with price or when the demand is elastic? With price elastic demand, scarcity pricing occurs (price rises above variable costs of the marginal plant) not only in a single hour but also on several other occasions. Further, if the demand for electricity decreases as prices increase then there will be a price above which prices cannot rise. Since prices cannot rise, the load/ demand will decrease till equilibrium is reached in the market.

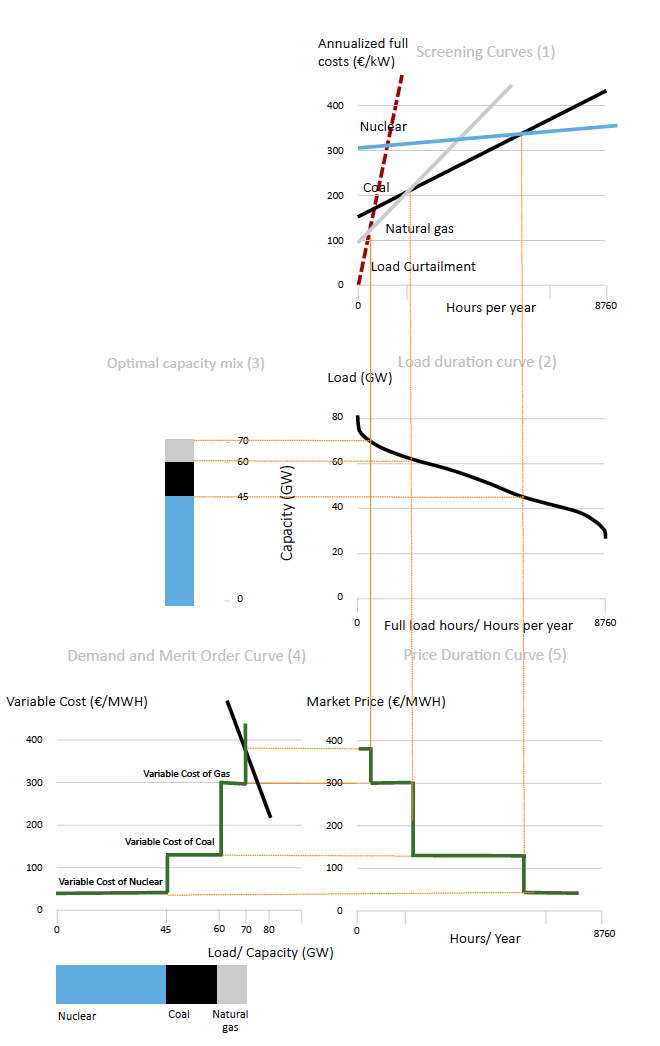
The equilibrium described above is not costless. The load curtailment required to balance demand and supply during peak hours will lead to loss of economic output/ utility which is termed as the “value of lost load” (VOLL).

**Graphical analysis**. Figure 11 (adopted from [Green (2005)](https://doi.org/10.1093/oxrep/gri004)), shows how scarcity pricing works. Panel (1) shows the usual screening curves along with a “load curtailment” curve that shows the value of lost load from not meeting demand. Note that for most of the hours during a year the loss from load curtailment is very high. But in some cases it is cheaper to curtail the load that to produce electricity at any cost.

Following the screening curves and the load duration curve we can derive the optimal capacity in Panel (3). This similar to the The optimal thermal generation mix derived earlier. The optimal installed capacity consists of 45 GW of nuclear, 15 GW of coal and 10 GW of natural gas i.e. a total of 70 GW. But the installed capacity is actually lower than the peak demand.

How does this long-term optimum capacity reflect in prices? Considering the 70 GW of installed capacity, we can derive the merit order curve or the supply curve of electricity. This is shown in Panel (4). The Panel also shows the demand curve for electricity during a scarcity period i.e. the demand exceeds supply. During this period, the price rises to above the variable cost of the marginal gas power plant. This will happen on all occasions during the year when demand exceeds 70 GW. This can also be seen in Panel (5). This is also the period during which the marginal power plant can recover its fixed costs. When demand is lower, the price will be equal to the variable cost of the marginal plant, which may be nuclear, coal or gas depending of the level of demand as is reflected in the price duration curve in Panel (5).

Figure 11. Peak-load/ Scarcity pricing and long-term equilibrium



**Source:** OEE, adopted from [Green (2005)](https://doi.org/10.1093/oxrep/gri004)

## Insights from the scarcity pricing model

**Market power**. The peak- load pricing theory shows that this set of prices is the only price set that is compatible with the long-term economic equilibrium, i.e. the conditions that investors recover their fixed costs, but do not make any long-term profits (nor losses). It does not, however, explain in detail how the price is determined in scarcity periods. It can only rise above variable cost, because the last generator can exercise market power: it is the only supplier left with sparse capacity. Obviously, it is challenging in this situation to distinguish “adequate” scarcity pricing from “inadequate” abuse of market power.

**Contestable markets**. Abuse of market power is inconsistent with a long-term economic equilibrium if market entry is possible. This is the idea of a “[contestable market](https://en.wikipedia.org/wiki/Contestable_market)”: if profits could be earned new generators would enter the market eating up all profits. Incumbents, anticipating this, behave like price takers in spite of having market power. The features of perfectly contestable markets – no entry barriers, no sunk costs, access to the same level of technology – are moderately well satisfied for peaking technology.

**Missing money**. If, for whatever reason, the electricity prices cannot rise high enough to recover fixed costs, as the peak-load pricing theory suggests, investors make a loss in the long run. This concern has been dubbed as “missing money problem”. This term has caused significance confusion in the past partly because observers who had not studied the idea of scarcity pricing also used it. But there are real reasons of concern, ranging from very practical issues (for example in Germany power exchange does not allow to enter prices above 9,999 €/MWh in its bidding form) to the fear that regulators might not be able to distinguish between scarcity pricing and abuse of market power. In the latter case regulators may face pressure to cap prices if these seem unreasonable. These concerns are plausible: any reasonable person would get suspicious if the price of something suddenly rises by a factor of 100 and some economists argue that “this is in line with theory”.

**Further reading.** Peak-load pricing theory was first published by Marcel Boiteux in 1949, an economist and mathematician that later headed Electricity de France (see [English reprint of 1960](http://www.jstor.org/stable/2351015)). The question received considerable attention in the 1950s and 1960s, with early independent contributions by [Houthakker (1951)](http://www.jstor.org/stable/2226608), [Steiner (1957)](http://www.jstor.org/stable/1885712) and [Hirshleifer (1958)](http://www.jstor.org/stable/1882234) and multiple generalizations thereafter. A useful survey of this literature is provided by [Crew et al. (1995)](https://link.springer.com/article/10.1007%2FBF01070807?LI=true) and an accessible discussion is included in [Green (2005)](https://doi.org/10.1093/oxrep/gri004).

**Other long-term price models**. The “screening curve model” introduced above is the simplest long run model of the electricity prices. Most models that are used in practice rest on the same reasoning, but include a whole set of additional constraints and features. Such models can be specified as a system of equations that is solved numerically by a computer. An example of such a numerical model is the European Electricity Market Model [EMMA](http://neon-energie.de/en/emma/).