

INVESTMENTS IN GENERATION AND TRANSMISSION

FRANÇOIS LEVÊQUE and GEERT BRUNEKREEFT

Electricity liberalization brought, among others, a profound change in the terms of investment in both generation and transmission. Decisions concerning the construction of new power plants, in particular the timing and the technology mix (i.e., the proportion of hydro electricity, nuclear, thermal, etc.) now depend on decentralised initiatives of investors and not on public authorities. As for transmission, which remained a monopoly, the reinforcement and expansion of high-tension power lines are no longer directly controlled by the generators. System operators have greater leeway for initiative. Depending on the specific case, they can sell transmission rights, submit investment programs to the regulatory authority, or invest as they see fit. In short, investments in an electricity system that is open to competition will no longer be coordinated by the same mechanisms as in the past. The planning that enabled a monopolistic and vertically integrated producer to adjust base and peak load capacities, as well as generation and transmission capacities, has been replaced by a series of decentralised decisions partly based on prices. Ideally, an optimal level of investment in the electricity system would involve joint optimisation of investments in generation and transmission. In fact, the goal is to minimise the cost of electricity to consumers. From an economic perspective, generation and transmission are complementary goods; if the price of one decreases, the quantity sold of the other increases. The mechanism underlying this phenomenon is simple: consumers are only sensitive to the total price of electricity, as they do not consume the generated electricity and the transmission service separately. Consequently, if the price of a KWh falls, *ceteris paribus*, they will consume more electricity and demand a greater quantity of the transmission service. Consequently, investments in generation and transmission complement each other.

Sometimes, however, investments in generation and transmission are substitutable. For example, in an isolated region with limited interconnection with the grid, a rise in local demand can be satisfied by either reinforcing the line or building a new power plant within the zone. If both investments occur simultaneously, then neither will be profitable. When both activities are combined within a single firm, joint optimisation of investments is deemed self-evident, since the stockholder or manager maximises overall profits. In an electricity system that is open to competition, the visible hand of the manager fails to ensure coordination between generation and transmission. Trans-

mission is separated from generation in one way or another (i.e., accounting, managerial, or legal unbundling) in order to ensure that rival generators have equitable terms of access to the grids.

This new situation opens the door to strategic behaviour on all sides. In order to provide for future investments in transmission, the transmission system operator (hereafter, TSO) must be informed of future investments in generation. Conversely, to plan these investments in generation, producers require forecasts of the TSO's future investments in the grid. In order to escape from this deadlock, one of these stakeholders must "draw first" by revealing its intentions and proceeding with the investment. However, the first to invest becomes hostage to the other, since it is impossible to move a power plant, or pylons, without forfeiting the bulk of their value. This is the classical economic problem of the hold-up occasioned by stranded costs. The upshot is generalised underinvestment: each party, knowing that it may be taken hostage *ex post*, reduces investments *ex ante*. Thus, we cannot apply the idealised rule for investment in transmission, which would have the system operator plan investment by optimising transmission and generation as a function of future demand and then lay the power lines in the hope that the market will induce generators to invest according to plan.¹

Nonetheless, it is necessary to avert such waste by finding some way to coordinate investments in generation and transmission. Various instruments, such as financial transmission rights and a zone-based rate structure for the grid, have been proposed in the recent economic literature.²

Let us examine how to optimise the utilisation and size of the grid when generating capacity is optimal, and how to optimise the utilisation and volume of generating capacity when the grid is optimal.

Aside from grid constraints, what obstacles must the market mechanism contend with so as to yield a socially efficient level of investment in generation, i.e., a level that satisfies the users' needs at the lowest cost?

The optimal investment in electricity generation is precisely determined by economic theory, which addresses both total capacity and its distribution amongst power plant types. These latter, in fact, differ both in terms of variable costs, which are usually linked to the price of fuel, and fixed costs, which essentially reflect expenditures on construction. For a nuclear power plant, the former are low and the latter very high; for a gas turbine this is inverted. Consequently, nuclear plants should be used throughout

¹ Notice that application of this idealised rule not only runs up against the opportunism of generators. It also assumes that the SO (or the competent regulator) acts in the public interest, is able to forecast future energy demand, and is able to precisely define the optimal level of capacity (i.e., the number, type, and location of plants) and the grid configuration that will satisfy that demand efficiently.

² See in particular: Hogan, W.W. (1992) "Contract Networks for Electric Power Transmission," *Journal of Regulatory Economics*, 4: 211–242. Joskow, P.L. and J.J. Tirole (2004a), "Merchant Transmission Investment," *Journal of Industrial Economics*, 2004. Chao, H.P. and S. Peck (1996) A market mechanism for Electric Power Transmission, *Journal of Regulatory Economics*, 10, 25–29.

the year to meet base-load requirements, while gas turbines should only be called on to meet peak-load demand at times of the year when there are spikes in demand.

Simple models of optimal levels of generating capacity generally considers two to three types of different plants. They permit to demonstrate how to identify the load-duration curve for the 8760 hours in a year and how to translate it into an hourly price curve. Naturally, the highest price is found when demand is greatest. As this demand exceeds available capacity, the equilibrium price is not set at the marginal cost of the last unit generated, but rather at a higher level equal to the marginal opportunity cost of consumption (i.e., above which the last consumer prefers to forgo rather than consume). The gap between these two marginal costs thus allows the peak-load plant that operates for the shortest period during the year to cover its costs. Notice that this result contradicts the conventional wisdom that the electricity market is incapable of ensuring that plants' fixed costs are being covered. This confusion arises from overly hastily equating the equilibrium price with the marginal cost of generation. In the presence of congestion, as during extreme peaks in this case, the shortage must be managed and resources allocated to those economic agents on whom the lack of access imposes the greatest cost.

Furthermore, economic theory predicts that if the peaking plant that is used least covers its total costs, and if the allocation among the various means of generation is efficient, then all other plants can cover their total costs.

The preceding economic model assumes that there is no uncertainty in terms of demand.³ However, consumers' reactions to price changes are very poorly understood. Except in the case of certain large consumers, who adjust their consumption to variations in the real-time prices on the spot market or accept compensation for forgone consumption, information on the price-sensitivity of demand is inadequate. Most consumers are not confronted with hourly, or even daily, fluctuations in the price of electricity. Their consumption is measured on a monthly or quarterly basis, and they are charged a rate per KWh that is independent of the hourly distribution of their consumption. Shielded thus from real-time price volatility, they have no need to hedge against the risk of high prices. Furthermore, most domestic consumers cannot be disconnected individually. And yet, there is no reason to believe that residents of residential neighbourhoods will face the same opportunity cost of not consuming. However, since they are all hooked into the same distribution network, creating a market of interruptible contracts cannot be envisaged. Consequently, there is no mechanism for revealing households' willingness to pay during peak hours.

Note that the underlying problem of short-term price-inelasticity of demand did not originate with the opening of electricity systems to competition. Under the previous arrangement, estimates of the value of electricity lost in the event of a service interruption (Value of Loss Load, or VOLL) were simulated by the planner in order

³ It also assumes risk-neutrality of investors. Risk aversion leads to under-investment in peaking plants – some of which are only profitable, in principle, if they operate several hours per year on average.

to decide when generation capacity needed to be boosted. When the cost of the new investment was lower than the benefit of the averted service interruption—VOLL multiplied by the reduction in risk of blackout attributable to the increased capacity (Loss of Load Probability, or LLOP)—, the investment was deemed worthwhile. To fix an order of magnitude, if VOLL is €10,000 per MWh, then the public interest is served by the construction of a power plant that will reduce the risk of interruption by approximately five hours over the course of a year. Today, with electricity systems that are open to competition, VOLL can also serve as a reference value. For example, during critical periods, a systems operator may decide to purchase power at a price equal to VOLL. In this event it is acting in the name of, and on behalf of, consumers.

However, it is quite unusual for the regulatory authorities to authorise such an astronomical price on the spot market, even during critical periods. The very potential of prices to reach that level provides a powerful incentive to generators to withdraw some capacity from the market so as to drive up the price – i.e., to exercise their market power during periods of tensions between supply and demand. Thus, for reasons of social acceptability and market power, the spot price is often capped by regulation at a level far below the VOLL. This type of intervention inevitably distorts the market signal towards under-investment, and the plant with the shortest period of operation during the year can no longer cover its fixed costs. The entire cascading structure for covering the fixed costs of the various plants collapses.

When real-time market prices are capped, undercutting investments, it becomes necessary to invoke other instruments to provide economic agents with a signal for the optimal capacity level. One elegant approach is based on the notion that generators do not supply a single good, electricity, but rather two goods, energy and capacity. The consumer values two services, the power itself when she wants to watch television or turn on a light, and also an option value for being able to do this at any time. From this perspective, generators should be compensated for the capacity they supply regardless of their utilisation.

Whether the selected system is obligation capacity or capacity payments, it is essential to bear in mind that the signals sent to investors originate at least as much from public authorities as from private agents. These are on the side of the invisible hand of the market all decentralised consumption and generation decisions that propel the evolution of the price. And they are on the side of the visible hand of public intervention the identification and the setting of the price cap and calibrating VOLL. We shall see this hybridisation recur in the case of investments in transmission.

Like other network infrastructures, electricity transmission grids present technical and economic characteristics that are quite challenging from the perspective of resource allocation. Like highways and airport runways, electrical transmission lines are congested. As a result, use of these infrastructures by one agent may degrade the quality of service available to another. In economic terms, this is known as a negative externality. In the case of electricity, congestion may even result in the complete col-

lapse of the system. If the electric current is not cut, the lines may stretch and melt! Again, like in the case of highway and airport infrastructures, investment occurs in discrete units, leading to discontinuous jumps in capacity. To expand a highway or an airport, a lane or a runway must be added in a single stroke. Smaller, fractional investments are impossible. In electricity, the line type for the high-voltage grid cannot be modulated by a single KV at a time. For example, either 220 or 400 KV must be chosen. Similarly, the gauge of the cable is not available in increments of a millimetre.

These two technical-economic characteristics, congestion and indivisibility (or lumpiness), are sometimes evoked in defence of misguided concepts. A first misconception is that investment must proceed until congestion is eliminated. In fact, if it were necessary to reinforce electricity transmission lines to the point that their capacities would be able to carry any and all transactions between generators and consumers at all times, the grid would be bloated and astronomically expensive. If, during a single hour in one year, a plant that is remote from a consumption zone is €10 per MWh cheaper than a local, more expensive generator, and if one MW of that generation cannot be transmitted to the consumers because of an inadequate line rating, then that line is congested during that hour. The cost of this congestion is €10 per year. Clearly, adding one MW of capacity to that line would be much more expensive! Eliminating all congestion would only make sense if grid construction costs were nil. Obviously, this is not the case, and consequently the economically optimal level of congestion is not zero. In fact, it is found at the point at which the cost of reinforcing the grid is equal, at the margin, to the savings it makes possible, i.e., electricity that can be bought from farther away at lower cost. The second misconception is that investment should be undertaken as soon as the new line construction project is profitable. It may, indeed, be preferable to wait and opt for a much more profitable project later, one which will add far greater capacity at a single stroke.

As with any infrastructure, it is worthwhile to distinguish between efficient use of the network and efficient size of the network. In the first case, capacity is treated as a given. Economic optimisation is thus a matter of allocating its use to the economic agents who value it most highly. Theory predicts that the key to accomplishing this lies in setting the access price equal to the short-term marginal cost. In electricity, this cost has two components. The first is due to ohmic losses that make it necessary to inject more electricity than is withdrawn at the other end of the line. The second component is congestion, which makes it impossible to replace local, high-cost electricity with less expensive power from a more distant plant. Notice that both of these elements of the marginal cost can be expressed as a function of the price of the transmitted power itself, for example in €/kWh. This allows us to establish an equivalence between the marginal cost of transmission and the marginal cost of generation. Between two local competitive markets, the equilibrium transmission price will equal the difference in marginal production costs, so that a buyer will be indifferent between buying from a seller who is closer but sells at a higher price and one who is farther away and sells more cheaply. The

energy pricing system that corresponds to setting electricity transmission fees equal to the short-term marginal cost is called nodal pricing, or marginal locational pricing. These terms reflect the fact that the electricity price is different at each node of the network. It also varies across time, given that demand, and by extension congestion, fluctuates between the nodes. For example, the systems operator of PJM (Pennsylvania-New Jersey-Maryland), the largest electricity market in the United States, computes the price at the 3,000 nodes several times per hour. The issue of efficient network size is an issue of optimal investment. The goal is to achieve the equilibrium size, i.e., expand capacity to the point at which marginal cost rises above the benefit yielded by the expanding of the capacity. In electricity, we have seen that this benefit amounts to displacing local generation with more remote, cheaper generation.

The distinction between efficient use and efficient investment arises because of a discrepancy between short-term and long-term marginal costs – the former being lower than the latter – and between marginal and average costs – the former again being the lower. These discrepancies are explicable in terms of contingencies, as well as by the presence of lumpiness and economies of scale.

The essential result of the realities described above and the discrepancies they give rise to is that a price equal to the short-term marginal cost ensures efficiency in use, but does not fully cover the investment expenditures necessary to construct an optimally sized grid. In other words, the nodal pricing system does not compensate the fixed costs of investments in transmission.

This is what keeps the market from operating efficiently. We note that, if there were no gap between the short-term marginal cost and the average cost, then a decentralised mechanism leading to an optimal level of investment might have been feasible. The underlying principle is to allocate transmission rights that yield congestion rents to the owners of each line as they are generated. Decentralised investments in transmission lines (called merchant lines by convention) are thus confined to modest growth. Thus, concludes Joskow,⁴ “*Most transmission investment projects are being developed today and will be developed in the future by regulated entities*”.

⁴ Patterns of transmission investments, Working Paper, March 2005. This paper is available at http://econ-www.mit.edu/faculty/index.htm?prof_id=pjoskow&type=paper.