Chapter 4

COST RECOVERY AND SHORT-RUN EFFICIENCY

Claude Crampes
Gremaq and Idei, University of Toulouse, France

When designing a tariff for the transport of electricity, the main difficulty is that the transport industry apparently incurs high fixed costs and no real variable cost. In effect, when the infrastructure is installed and when the operators are at their workplace, the only input which is necessary to deliver electricity at a given withdrawal node is electricity at some injection node since electricity is flowing by itself. Consequently, at first sight the problem is just to allocate fixed costs, mainly infrastructure maintenance costs, wages and financial charges, among the different types of users of the grid.

As a matter of fact, the transport of electricity creates two significant variable costs.[1] One is an internal cost, that is, a cost in terms of electricity: a fraction of the energy which is injected into the grid will be lost during transport. It results that the consumption of 1 MWh of electricity requires the generation of $(1+L)$ MWh, and this extra L MWh is a real cost for which producers must be compensated. Additionally, because the lines and nodes used for transport have a limited capacity, the optimal allocation of production and consumption is not as efficient as it would be, absent any grid constraint. And pricing must include this economic cost due to congestion.

When transport prices are computed using the marginal values of these two costs (the so-called "nodal prices"), they provide a revenue larger than the losses of energy. This surplus can be used to pay for fixed costs of transport but, in most cases, it is not large enough to balance the budget of the operator. This explains why transport tariffs must be either second-best linear prices (Ramsey prices) or non linear prices.

The chapter presents the normative principles of the economic analysis of pricing in electricity transport. It is exclusively dedicated to short-term

analysis, that is to the operation of a given electricity network. The development of the transport infrastructure is analysed in Chapter 5.

In Section 1, we define the variable cost of transport as the sum of the cost of congestion and the cost of ohmic losses. In Section 2, we show how nodal prices can be used to design tariffs for transport. Section 3 is dedicated to the problem of fund rising in order to balance the budget of the transport operator. Firstly, we focus on Ramsey prices, which are the second-best linear prices when one tries to reach efficiency without impairing the budget equilibrium of the transport firm. Secondly, we consider a special class of non-linear prices, namely, two-part tariffs. We conclude in Section 4.

We do not discuss how to define transmission rights on the transport infrastructure and the conflict between the supporters of physical rights and the supporters of financial rights.[2] Neither do we consider the ownership and the governance of the transport firm. We suppose that users face no barrier to gain access to the grid. The regulation of the transport operator under alternative hypotheses concerning vertical integration and competition between grid users is scrutinised in Chapter 6.

While in the following sections we have adopted a non-technical presentation, the reader can find in the appendix a formal modelling of the main results.

1. FIRST-BEST DISPATCH IN AN ELECTRICITY NETWORK

1.1 The transport of electricity

In economic terms, a good is defined by: *(i)* some intrinsic characteristics (weight, size, quality, etc.); *(ii)* the location; *(iii)* the date; and *(iv)* the state of nature where it is available. Transport is the activity that mainly consists in modifying attribute *(ii)*, even if, as side effects, the three other attributes are also modified in most cases. It results that, to analyse the utility of transporting a specific good, we need to analyse the utility and the cost of that good at the departure and arrival locations. The difference between the net utility of the good at the arrival location and at the departure location is the gross utility from transporting it. This difference is to be compared with the cost of transport in order to decide if the good is to be displaced or if it should remain at the initial location.

This is the normative principle that we have to apply when analysing electricity transport. The starting point is to determine the quantities to

generate and to consume in order to maximise the welfare of all the agents that use the transport infrastructure. In the short run, the equipment for generation, transportation and distribution is fixed. The preferences of consumers also are fixed. The optimal allocation is limited to deciding how much to generate at each node and how much to consume at each node, given the restraints imposed by the topological characteristics of the network and the technical capability of each piece of equipment (see Box 4-1 for details on the objective to maximise welfare under alternative sets of constraints).

Box 4-1: First-best, second-best, and constraints to maximise welfare

In the short run, the whole generation and transportation equipments are fixed, as well as all the needs.

- The "grid-free" first best allocation is the set of quantities of electricity generated and consumed at each node that maximises the net welfare, that is, the sum of the difference between the utility of electricity for consumers and the cost to generate it, in a fictitious situation where energy can flow from one node to others without any constraint and without any loss.

- The "grid-constrained" first best allocation also maximises the net welfare, but taking into account the physical characteristics of the grid. In this case: *(i)* some energy is lost during transport; and *(ii)* because some lines and intermediary nodes have limited capacity, the "grid-free" optimal flows are no longer feasible, which creates a "congestion cost".

- In economic theory, "second-best" mainly refers to a situation where the benevolent planner has to balance the budget of the producers he supervises. Actually, this expression can be used in any situation where a constraint is added to an initial allocation problem. In that respect, the "grid-constrained" first-best is a second-best with respect to the "grid-free" first best allocation.

- When a constraint is added to a given optimisation problem, either that constraint is not binding and the initial allocation does not change, or it is binding and it results in a decrease of the initial performance. The new allocation can never give a higher performance since, if feasible now, it was feasible before the new constraint is added and it would have been chosen. The difference between the performance without and the performance with the constraint is the economic cost of the constraint. For example, the economic cost of the transport grid is the difference between the grid-free social welfare and the grid-constrained social welfare.

- In the same way, one can measure the cost of additional constraints, such as:
  – the obligation to balance budget;

– the prohibition to discriminate on prices;
– restrictions on tariff classes (linear, two-part, etc.);
– the inability of the operator to collect information on preferences and costs;
– the universal service obligation;
– etc.

- Short-run decisions are constrained by the incapacity to adapt the transport infrastructure and the generation plants. For this reason, the short-run cost, in which there are significant fixed costs, is higher than the long-run cost exclusively made of optimally chosen variable inputs.

To understand why the optimal allocation can be defined in that simple way, it is interesting to stress some important differences between the transport of electricity and other transport activities, for example freight or passengers transport. Firstly, electricity is highly standardised, which means close substitutability between generation nodes for a given need and, symmetrically, close substitutability between consumption nodes for a given production. This homogeneity property allows to pool quantities. Secondly, for electricity, the time attribute of the good is not modified by transport. Power flows instantaneously through lines and intermediary nodes. (See Box 4-2 for some illustrations of network topologies). This explains why, in the optimal dispatch, injections and withdrawals are contemporaneous. For the same reason, the dispatcher can know for sure the state of nature at a withdrawal node when injecting power at another location, which greatly reduces the randomness of net locational utilities. The third difference with most transport networks is that, for electricity, the technology does not allow to control the physical flow of energy through the grid (see Box 4-2 for an illustration). Consequently, the actual flow on each line cannot be a control variable. Fourthly, in a given grid, one can predict very precisely the amount of transport losses because they follow well known physical laws.

To sum up, transporting electricity consists in controlling modifications in its attributes *(ii)* and *(iv)* without modifying attribute *(iii)* and provoking an undesirable but predictable change in attribute *(i)*. The physical path followed because of the transformation in attribute *(ii)* cannot be controlled; it results that transport between two nodes in a meshed network creates externalities on all lines and nodes intentionally and unintentionally crossed by the energy flow.

In the optimal allocation defined by a benevolent social planner, at each node, marginal utility is equal to marginal cost. If, at one node, marginal cost were higher than marginal utility, the last kWh would be generated at loss and, symmetrically, if marginal cost were less than marginal utility, it would mean that the entire potential social surplus is not created. At each

node, this equality of marginal cost and marginal utility will most likely require a transfer of energy. Some nodes will have to export energy and others will be net importers, depending on the cost structure of generation and given the consumers' preferences for electricity. When no piece of infrastructure exhibits congestion, and when there is no ohmic loss, the optimal allocation is such that energy has one single value throughout the whole network, which can be viewed as a giant unique node or as a plate. To see this, observe that if there remained a difference between two nodes, it would be easy to increase the global surplus by transferring some kWh from the node with the lower marginal valuation towards the node with the higher valuation.

Box 4-2: Of nodes and lines

An electric grid can be viewed as a set of nodes, either final (injection and withdrawal nodes) or intermediary (transformers, meters, controllers, etc.), interconnected by lines.

The simplest network is made of one single line connecting two final nodes. The "north-south" network represented below (see Figure 4-1) is a useful theoretical configuration to understand congestion and losses, but it also gives a reasonably good picture of the grid in some countries.[3] Since there exists a single line, there is one unique possible path for transporting electricity from north to south or from south to north. Absent any energy loss, the physical equilibrium of the electric industry imposes $q_n^g + q_s^g = q_n^w + q_s^w$, where q_n^g (respectively q_n^w) stands for the quantity generated (respectively consumed) at the north node and q_s^g (respectively q_s^w) stands for the quantity generated (respectively consumed) at the south node. Consequently, the quantity of electricity flowing on the line is $\left|q_n^g - q_n^w\right| = \left|q_s^w - q_s^g\right|$.

But in many countries, particularly in continental Europe, electric networks are meshed. The consequence is that there is not one unique path for electricity to go from one node to another. This is illustrated in the three-node network hereafter (see Figure 4-2). Energy flows follow paths of least resistance determined by Kirchhoff's laws. Suppose a generation node and a consumption node are connected by two lines, one with a resistance twice the other's. When generators inject quantity *q* at one node and, assuming no losses, consumers withdraw the same quantity at the other node, the flows on the low resistance line and on the high resistance line are respectively *2q/3* and *q/3*. In a 3-line network with the same resistance on each line, like in Figure 4-2, if there are two generators installed at nodes *1* and *2* respectively, and consumers are located at the third node, the simultaneous injection of q_1^g and q_2^g will generate a superposition of flows on the lines connecting consumers to generators. For example, the line between node 1 and node 3 transports two thirds of the energy injected at node 1 plus one third of the energy injected at node 2. By contrast, only one third of the *net* flow circulates on the line between the two generators.

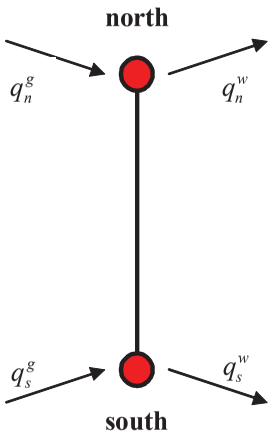

Figure 4-1. One line network

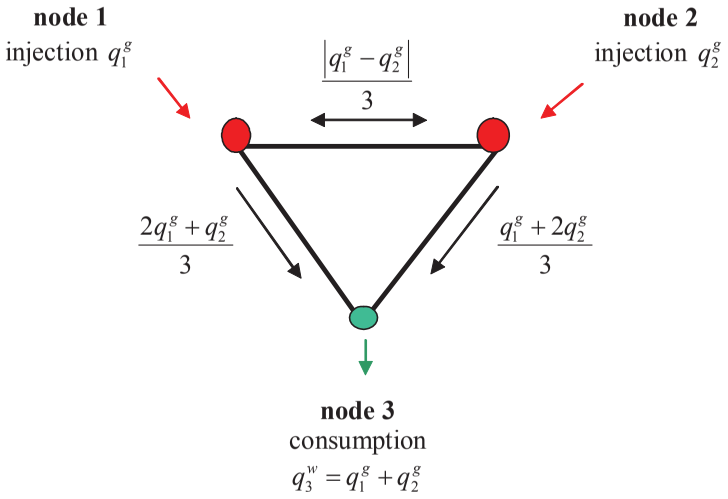

Figure 4-2. Three-line network

But, as explained below, because of losses and because of some scarce capacity in transport, transfers from one node to another cannot be done for

free. It means that the "very-first-best" allocation (or "grid-free" allocation) is not feasible. The dispatcher can only reach the "grid-constrained" optimal allocation and, consequently, energy valuation resulting from this dispatch will differ among the nodes.

1.2 The cost of congestion

Assume first that losses can be neglected. At importing nodes, the incoming energy is limited by the capacity of lines and transformers along the physical path followed by energy. As a result, electricity is relatively scarce and it is more valued than if there were no capacity constraint. Reciprocally, at exporting nodes, energy is relatively in excess and low valued because the outgoing energy is limited by the capacity of lines and transformers. The more congested the network, the higher the discrepancy between nodal valuations. The limit case is the autarky case where lines are cut sot that each node is isolated from the others.

The difference between two nodal valuations of energy, absent any loss, is an index of the tightness of the transport constraint. It reflects the incapacity of the operator to increase generation at low-cost nodes and to decrease it at high-cost nodes as well as its incapacity to increase consumption at high-utility nodes and to decrease it at low-utility nodes. The "merit order" commands that no generator should be dispatched if there remains some available capacity with a lower cost. It is no longer implementable. In the simplest case of Box 4-3, with one single line connecting efficient northern generators with a south node where there are inefficient generators and the load, one can easily draw the grid-constrained optimal quantities and measure how they depart from the grid-free optimal quantities. The difference in nodal valuations exactly reflects the social cost of having an out-of-merit-order dispatch because of the limited capacity for transport between north and south. The difference is the shadow value of the transport line that signals by how much social welfare would be increased if the constraint could be relaxed.

In meshed networks, energy flows along least resistance paths without the possibility to control them.[4] As a result, any injection and withdrawal of a given quantity at two distinct nodes will provoke a transit of electricity through all the lines of the network. If one line is congested, all paths will appear congested. Consequently, in a meshed network, congestion on one line is sufficient for nodal values to differ throughout the network. Because of this "contagion effect", the dual value of the congested line is larger than the mere difference between the marginal values at the two ends of the congested line.[5] The difference between these two values reflects the

negative externalities in electricity transport, that is, the additional congestion cost due to loop-flows.

Box 4-3: Out-of-merit-order optimal dispatch

Consider the north-south network in Box 4-2 (see Figure 4-1). Assume that all consumers are in 'south' and their marginal utility from withdrawing the quantity q_s^w of electricity is given by the decreasing function $U'(q_s^w) = 10 - q_s^w$ (see Figure 4-3). In 'north', electricity can be generated at a constant marginal cost equal to $1¢$ and, in 'south', there exist plants with a constant marginal cost equal to $4¢$. There is no constraint of capacity for generation and there is no energy loss on the line. Let K denote the capacity of the line.

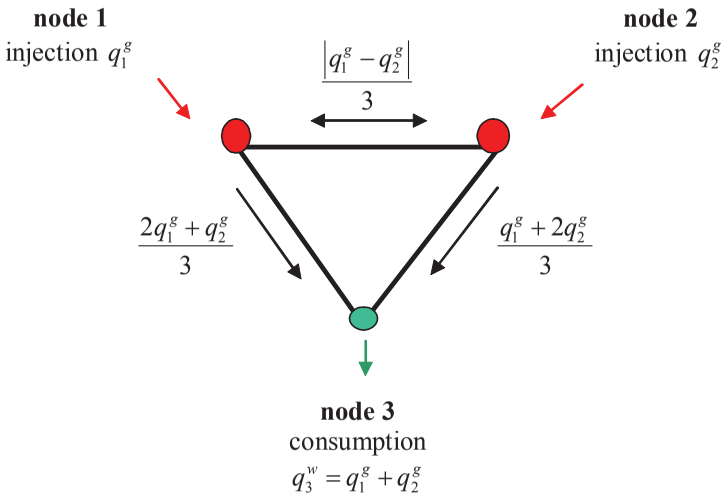

Figure 4-3. Out-of-merit-order optimal dispatch

If K is very large, the optimal dispatch consists in producing nothing in 'south' where generation is very costly. The whole energy comes from 'north'. The optimal consumption is such that $U'(q) = 1¢$, that is the grid-free optimal quantity q_s^w=9.

If $K < 9$, the capacity of the line does not allow to import this quantity of energy from 'north'. The dispatch first consists in saturating the line to transfer as much energy as possible from 'north', which is K. After that, there are two possibilities: *(i)* if the marginal utility from consuming K is still higher than the marginal cost of generation in 'south', use the south plant up to the point where the marginal utility of electricity is equal to the marginal cost of generation, that is $U'(K + q_s^g) = 4¢$. Otherwise, do not dispatch the costly plant. In the first case, the grid-constrained optimal allocation is $q_n^g = K$, $q_s^g = 6 - K$ (as long as $K \le 6$) and the total consumption is $q_s^w = 6$.

In the second case, that is for K between 6 and 9, the total output is the constrained flow coming from 'north', $q_n^g = K$, $q_s^g = 0$. These quantities are graphed in the middle panel of Figure 4-3 as functions of the capacity of the line K.

Because of the inefficient dispatch created by the limited capacity of the line, welfare cannot be as high as it would be if low-cost generators were located at the south node with consumers. The lower panel of Figure 4-3 represents the marginal value of the capacity constraint, which represents the marginal cost of congestion η. When K is smaller than 6, one additional unit of capacity would allow to substitute one north kWh to one south kWh, that is to save $\eta = 4¢ - 1¢ = 3¢$ for an unchanged total output. When K is between 6 and 9, one additional unit of capacity allows to increase the consumption by means of more imports from 'north', so that its value is the net marginal utility of electricity; $\eta = U'(K) - 1¢ = 9 - K$. Finally, when K is larger than 9, any development would be useless, which is signalled by $\eta = 0$.

As shown in Chapter 5, the shadow price of the lines is to be compared with the real price of one unit of equipment to know whether the transport capacity is to be increased or decreased. When the dual value of the constraint is higher than the cost of one additional unit of equipment, the transport line should be developed. And it should be downsized in the opposite case. In actual networks, the infrastructure is almost always larger than its optimal size. The consequence is that the shadow price of lines and transformers is less than their unit cost. And it is even equal to zero when there is no congestion at all.

1.3 The cost of losses

Suppose now that there is no congestion. The main cost of delivering 1 MWh at one node starting from a specific injection node results from the fact that a quantity L of the energy injected will be lost in transport. It means that $1+L$ MWh are to be generated. As a result, in an optimised network the marginal valuation of 1 MWh will differ from one node to the other by the value of the lost energy. Note that it is not a cost incurred by the transport grid itself. It is a cost due to the distance between injection and withdrawal nodes. It directly concerns generators and consumers.

In electric networks, losses increase proportionally to the square of the energy injected. The consequence of this precise functional form is that marginal losses are twice higher than average losses.[6]

1.4 The short-run marginal cost of transport

To sum up, because maintaining and developing the infrastructure is costly, it is optimal to keep some congestion in most pieces of the transport grid. As a result, the dispatch that would maximise net social welfare without any reference to the grid (like if all generators and consumers were located at the same place) is not feasible. The actual dispatch is sub-optimal as compared with the fictitious one-node industry. The cost due to congestion is equal to the difference between the maximum welfare obtained without transport constraints and the welfare that results from the actual dispatch. In addition to congestion costs, only a fraction of the quantity injected can be consumed. The cumulative effects of these elements is that in a network built for electricity transport, the optimal allocation of quantities to generate and to withdraw at each node is such that marginal valuations will be different from one node to others. The nodal valuation of energy is a natural by-product of the optimisation algorithms used by system operators. When an Independent System Operator computes the feasibility of a given dispatch on the grid he controls, the value of energy at each node can be published instantaneously. The difference between nodal valuations that includes a real cost (the value of lost energy) and a shadow cost (the value of lost efficiency), is to be viewed as the short-run marginal cost of transport. Yet, note that none of the two components of the short-run marginal cost can be directly related to an economic or accounting expenditure incurred by the operator of the transport infrastructure.

1.5 Time variation

The needs for electricity are strongly variable in time. They are both cyclical (according to well known daily, weekly and yearly variations) and random (for example because of changes in temperature). There occur other types of time variability on the generation side, due to the scarcity of hydro resources, the fluctuations of fuel prices and the availability of plants (maintenance and repairing). In contrast, the transport infrastructure is almost fixed for the medium run. The optimisation of a strongly variable welfare function under invariable transport constraints obviously results in a continuously varying optimal allocation and, consequently, a continuously varying marginal value of energy at each node. During low activity periods, the capacity of the lines and transformers is not binding. The nodal valuations differ only by the marginal value of losses, which are rather low since the level of consumption is low. In contrast, during peak periods, the difference between nodal valuations is very high. Note that the last assertion is not always true. In some circumstances, the difference between peak load

and off-peak load can be larger at an exporting node than at an importing node. In that case the electricity flowing on the line at peak periods can be smaller than at off-peak periods and the nodal difference in electricity valuation can be higher at off-peak periods.

2. NODAL PRICES

2.1 Energy prices

The first-best allocation can be decentralised by means of prices which reflect the marginal value of electricity. It means that the first-best generation levels and consumption levels would be freely chosen by individual generators and individual consumers if they were facing prices equal to the marginal value of electricity in the optimal allocation. And since, as shown formerly, marginal valuations change from one node to the other (and from time to time), the decentralisation of first-best necessitates nodal prices. If perfect competition mechanisms prevail at each node (roughly said, if there exists a large number of small buyers and suppliers of electricity at each node who behave as price-takers), the equilibrium will naturally determine energy prices equal to the marginal values obtained formerly. Consequently, organizing competition is a good way to reach efficiency. However, even when the number of generators and the number consumers are reasonably large, the decentralisation of decisions requires some public intervention, for example to organize the matching of demand and supply in wholesale markets.

In contrast, if there is only a small number of large suppliers and/or buyers, the equilibrium price would reflect the market power of these agents.[7] To implement the first-best, nodal prices are to be regulated by the government or, more precisely, by a regulation entity or an antitrust authority.

In any case, when consumers at node i can buy electricity anywhere and pay electricity coming from any node j at a price p_i equal to the optimal marginal valuation of node i, they consume the optimal quantity. And when generators at node i are authorised to sell electricity anywhere and receive p_i for any kWh sold to any node j, they produce the optimal quantity.[8]

2.2 Transport prices

For a given transaction, when the buyer and the seller are located at the same node i, they transact at the same price p_i that covers only generation

costs. If the seller is at node i and the buyer at node j, the former receives p_i and the latter pays p_j. The problem is to know what to do with the difference.

To give an answer, note that another way to decentralise the first-best is to distinguish the price of energy and the price of transport. When there are two separate bills, the price for transporting 1 MWh from node i to node j is to be $t_{ij=} p_j - p_i$. As a matter of fact, the consumer at node j must be indifferent between buying its energy at node j at price p_j on the one hand and, on the other hand, buying it at node i at price p_i and then paying t_{ij} for its transport to node j. Similarly, the generator at node i must be indifferent between selling its energy at node i at price p_i on one hand and, on the other hand, selling it at node j at price p_j while paying t_{ij} for its transport to node j. Consequently, we can conclude that, from the point of view of consumers and producers, the difference $t_{ij=} p_j - p_i$ really appears as a transport fee. For this reason, it is natural to pay it to the operator of the transport infrastructure even if the costs covered are on the users' side.

Note that if the net flow of energy is from i (the exporting node) to j (the importing node), any individual transaction from i to j increases congestion and must be charged $t_{ij} > 0$. In contrast, assume $p_j < p_i$ so that node j is a net exporter to node i and there occurs an individual transaction between a generator located at i and a consumer located at j. How can it be possible since the generator receives more than what is paid by the consumer? Because the individual transaction creates a counter-flow, it alleviates the congestion on transport lines. Consequently, the transport pricing system should promote this type of transaction by rewarding the parties instead of charging them for transport. And this actually occurs with nodal prices since the transport of 1MWh from i to j would then be charged $t_{ij=} p_j - p_i < 0$. In other words, the counter-flow transactions would be encouraged.

The total value of the flows of energy using the transport prices derived from nodal prices is the "merchandizing surplus". Even if counter-flows are to be rewarded, the merchandizing surplus is obviously positive[9] for the following reason. Recall that nodal prices reflect optimal valuations of energy. Consequently, the merchandizing surplus is made of two elements. The first piece is a net surplus due to the specific functional form of the losses of energy and their coverage by a marginal pricing rule. Actually, the generator must be paid for the electricity he produces, even if it cannot be consumed. But because losses increase with the square of the load, their marginal value is twice their average value. Consequently, after compensating the generator for its losses, it remains an equal sum that can be used to pay for transport. The second part is the cost of congestion, and its overall value cannot be negative. Either some lines or nodes are congested and then the surplus is positive, or there is no congestion and then the

surplus is zero. By its very definition, the merchandising surplus does not correspond to the payment of costs incurred by the operator.

The transport prices calculated using nodal prices vary with the date, the initial node and the terminal node of the transaction. All the users who transact at the same date, the same initial node and the same terminal node should pay (or should be paid for counter-flows) the same unit price. Therefore, we can assert that the nodal price system is objectively non-discriminatory (see Chapter 1).

3. CONSTRAINED AND UNCONSTRAINED PRICING

3.1 Efficiency concern *vs.* fund-raising concern

First-best pricing allows paying for all the costs of an industry when the equipment is optimally designed and there are no increasing returns to scale in the long run. In this hypothetical situation, on the one hand the short-run marginal cost and the long-run marginal cost are equal and, on the other hand, the long-run marginal cost is at least equal to the long-run average cost. Consequently, a price equal to marginal cost cannot be below the long-run average cost.

But these optimal conditions are practically impossible to meet in actual networks, because of the following reasons:[10]

(i) discrepancy between long-run and short-run needs

The optimal dynamic expansion of the transport infrastructure does not amount to the mere connection of many static plans. Most network facilities are built for 20 or 30 years and it is unlikely for any one of the facilities to have exactly the optimal static capacity for a given year. Additionally, deviations from the optimal plan are due to errors in the forecast of demand and generation costs.

(ii) technical non convexities

For transport facilities, decisions are discrete options rather than continuous variables. For a given reinforcement, the feasible set is very small, *e.g.*, lines of 220 kV or 400 kV. Consequently, as compared with the "marginal optimum", the result of programming with integer number will be an apparently either oversized or undersized network, with a natural tendency towards oversizing when demand is expected to increase.

The tendency towards over investment is intensified because of scale economies: the larger the capacity of the investment the smaller the capital cost per unit of capacity.

(iii) additional constraints

Network expansion planning is subject to reliability constraints which create the need for extra capacity with respect to the optimal plan under strict economic terms. Financial, environmental, technical or even political restrictions are also imposed on the expansion of the network in general or of a given corridor.

We conclude that, because the infrastructure is oversized as compared with short-run needs and because there exist long-run economies of scale, congestion rents calculated on the basis of short-run marginal costs will be rather low, and the rents from losses will not be high enough to cover all the fixed costs incurred by the grid operator.

3.2 Budget constraint

Several solutions are available to balance the budget of the transport firm. If public subsidies are allowed, they can be used to pay the difference between the merchandising surplus and the fixed costs of the infrastructure. But the taxes levied to fund the subsidies create economic distortions and the resulting allocation of consumption and production cannot be the first-best. Because of the cost of public funds, we can only reach a second best. Additionally, within the European liberalised framework, the State-aid solution is not authorised. Therefore, prices are to be adjusted to cover all the costs.

When constrained to balance the budget of the transport operator, the optimal tariffs will depart from first-best if specific restrictions are imposed in their computation. When any functional form can be implemented, that is when the bill is not necessarily proportional to the quantity of energy that is transported (non linear prices), the first-best allocation remains feasible.[11] Among non-linear tariffs we will only discuss two-part prices. On the contrary, if prices are restricted to be linear, the best ones are Ramsey prices, and the resulting allocation of generation and consumption will only be a second best. We first consider the case where only linear prices are feasible.

3.3 Second-best linear prices

Ramsey prices are the best linear prices when the balancing of budget is mandatory. They cover marginal cost (congestion and losses) plus a margin computed to pay for the fixed costs. In methods for pricing applied by practitioners, fixed costs are allocated according to some proportionality rule

(for example the cost share paid by agent *i* when consuming the quantity q_i is:

$$q_i / \sum_j q_j \text{ , or } C_i(q_i) / \sum_j C_j(q_j) \text{ ,}$$

where $C_i(q_i)$ is the direct cost to produce q_i, or any other ratio[12]). In Ramsey prices, one takes into account the reaction of the infrastructure users when they are billed or paid. Pricing above marginal cost is inefficient because it provokes a decrease in consumption. Similarly, when a producer is paid less than marginal utility, he decreases his production. Consequently, the aim of second-best pricing is to limit this bias in quantities produced and consumed. It is easy to understand that, if some consumers react to price increases less than others, it is optimal to make them pay more than more reactive consumers. And symmetrically, if some generators decrease their output more than other producers when the selling price is decreased, it is optimal to modify their selling price less. This explains that Ramsey prices are inversely related to the elasticity of demand and to the elasticity of supply.

As a consequence, they are discriminatory: each segment of users that can be isolated from the others will pay or will be paid a different Ramsey nodal price. In the simple case where, at each node, one cannot distinguish between different types of consumers and between different types of producers, it remains true that one can distinguish a group of consumers on one side and a group of producers one the other side. Consequently the Ramsey demand price p_i^w will be different from the Ramsey supply price p_i^g: because of the need to levy funds to cover fixed costs, we lose the property of price uniqueness at one node.

When energy and transport of energy from *i* to *j* are billed separately, the equilibrium condition is that paying p_j^w to consume at node *j* is to be equivalent to buying at price p_i^g at node *i* and transporting towards *j*, that is $p_j^w = p_i^g + t_{ij}$. It results that the Ramsey price for transport $t_{ij} = p_j^w - p_i^g$ is the weighted sum of two elements: *(i)* the cost in terms of efficiency due to energy losses and congestion we already had in Section 1; and *(ii)* the cost in terms of efficiency due to the budget constraint. The first part can be positive or negative depending on the direction of individual operations as compared with the net flow between *i* and *j*. The second part is necessarily positive: it depends on the elasticity of the energy supply function at the injection node and on the elasticity of the energy demand function at the withdrawal node. The smaller these elasticities, the higher the Ramsey price for transport between *i* and *j,* whatever the costs of ohmic losses and congestion.

Note that Ramsey pricing commands a global allocation of fixed costs based on demand elasticity, but transparency and political arguments impose some fragmentation of fixed costs. Actually, the obligation to balance the budget can be solved in many different ways. One possibility is to impose a global budget constraint to the transport operator. In this case, Ramsey prices allocate all the fixed costs of infrastructure to all users. Lines and transformers in low voltage are paid by all users, including large consumers and generators who use only high or medium voltage equipment. Another solution is to exempt large users from paying for the fixed cost of the low voltage system. In this case, the optimal allocation of consumption and generation among nodes and the resulting flows on lines is to be calculated subject to several budget constraints. As explained in Box 4-1, the larger the number of constraints, the higher the efficiency loss. It means that, when price discrimination is allowed, discomposing the cost recovery constraint into several user-targeted constraints cannot be justified in terms of efficiency.

3.4 Two-part tariffs

The drawbacks of Ramsey prices are obvious: *(i)* a large quantity of information is necessary to compute the elasticities; *(ii)* discrimination is forbidden by law, *(iii)* the resulting quantities deviate from first best. For these reasons, multi-part prices can be preferred.

In the simple case of two-part prices, the consumer pays a fixed amount of money independent of the quantity of transport he will require and, then, he pays a "marginal price" for each unit he wants to transport. Since the marginal price can be set equal to the marginal cost of congestion and losses, this tariff allows implementing the first-best allocation, provided that the fixed part of the tariff does not exclude any user with a marginal willingness to pay higher than the marginal cost. In effect, the risk with one unique two-part price is that the fixed fee must be high enough to cover all the fixed costs, so that users with low income will not be able to pay for it. If these low-income users can be easily identified, a solution is to propose them a specific two-part price, for instance with a zero fixed part. The flaw is that we are back to discrimination. Even a single two-part price is discriminatory since the unit price paid is decreasing with the quantity consumed. Insofar as two-part prices are acceptable on legal grounds, it is efficient to propose to users not one but a whole set of two-part prices, letting each user choose within the menu the tariff he prefers. It is a very efficient way to collect funds for fixed costs covering. It is also a good way to organize "second degree price discrimination" when the price maker lacks information about the willingness to pay of users.[13]

Applying these principles to the transport of energy, the tariff to transport a given quantity form node *i* to node *j* should have a first part proportional to the quantity (where the proportionality coefficient is the marginal cost of congestion and losses) and a second part, fixed, computed after the fixed parts of the energy tariffs for consumers at node *j* and generators at node *i*. The fixed part of the tariffs for energy must be adjusted not only to pay for the fixed costs of generators but also to pay for the fixed costs of the grid operator.

By definition, the fixed part of two-part tariffs does not distort efficiency. Therefore, the tariff maker has some degrees of freedom to calculate them. He mainly has to avoid the exclusion of some agents, individually or collectively. In electricity transport, the main concern is that large consumers have alternative opportunities to the use of the grid, for example by installing generation plants at their location. To prevent this by-pass, several two-part tariffs must be proposed. Each item in the menu is tailored so that no one is excluded (participation constraint) and each agent chooses the tariff calculated for him (self-selection constraint). When also taking into account the possibility of coalition between several producers or large consumers, or between a producer and one of its clients in order to bypass the transport infrastructure, it is necessary to have recourse to cooperative-games theory to optimally design the fixed part of the tariff.[14]

Finally, note that devices able to continuously meter the flow of electricity on a line are very expensive and, in many countries, they are installed only on the high voltage grid. Additionally, when available, it would be very costly both at the operator level and at the user level to process all the information they collect. As a result, instead of sending invoices based on the time profile of the demand for transport, in most countries one uses to calculate the aggregate quantity transported during a given period (for example, one year) and to charge a uniform marginal price of transport all the year long. The drawback is that the same total quantity of energy can be transferred with strong time regularity or with a high irregularity. And the irregular load profile is much more demanding in terms of transport capacity than the regular one. One solution is to distinguish an energy component and a capacity component in the demand for transport, each with a specific constant marginal price. But it remains true that these two elements are to be designed to send accurate signals of scarcity to users. A fixed part remains necessary to cover the fixed costs of the infrastructure without distorting the decision of the users.

4. CONCLUDING COMMENTS

To calculate the tariffs to transport electricity, the trade-off between efficiency, budget requirements, legal constraints and practicability, results in a non linear tariff based on the nodal prices of energy and on the fixed costs of the infrastructure. The simplest is two-part. The total bill to be paid to the transport firm is made of an "efficiency-oriented" part, variable with the quantity transported, and a "cost-recovery" part, independent of the quantity transported.

The first part is aimed at signalling the cost of physical losses and the cost of congestion. It should vary with time and location and should be increasing with the quantity of energy transported.[15] Practically, it should be a function of the injection and withdrawal nodes (not on the physical path of energy since it actually cannot be controlled) and on the direction of the flow. For practical reasons, this first part of the tariff can be discomposed into an energy term and a capacity term, both linear functions, provided that the scarcity signals they transmitted are similar to the ones that allow an efficient use of the grid.

The fixed part of the tariff is aimed at paying for the infrastructure costs. Because its purpose is purely budgetary, it should not interfere with the scarcity signals sent by the variable part(s). In particular, when the transport tariff distinguishes between an energy component and a capacity component, the latter should not be viewed as the fixed part of the tariff.

In most countries, actual transport tariffs are two-part but the way they balance the variable part(s) and the fixed part is very heterogeneous.[16]

APPENDIX

MODEL SETTING

Two nodes, *n* (for north) and *s* (for south), are connected by a line of capacity *K*. The unit cost of capacity is *r*. The quantity of electricity transported through the line is measured in the same unit as *K*. We note $U_i\left(q_i\right)$ the utility derived from the consumption of a quantity q_i of electricity at node *i* and $C_i\left(q_i\right)$ the cost of generating q_i at node i $(i = n, s)$.

Let $W\left(q_n^w, q_n^g, q_s^w, q_s^g, K\right) = U_n\left(q_n^w\right) - C_n\left(q_n^g\right) + U_s\left(q_s^w\right) - C_s\left(q_s^g\right) - rK$ stand for the welfare function of this industry, where superscript *w* stands for "withdrawal" and superscript *g* stands for "generation". We consider only short-run decisions, that is, *K* is not a control variable. The optimal allocation is the solution to:

$$\max_{q_n^g, q_n^w, q_s^g, q_s^w} \quad W\left(q_n^w, q_n^g, q_s^w, q_s^g, K\right) \tag{1}$$

s.t. $$q_n^g - q_n^w = q_s^w - q_s^g \tag{2}$$

$$z_{ns} \overset{def}{=} q_n^g - q_n^w \leq K \tag{3}$$

Provisionally, we assume there is no loss, which is reflected by (2). To simplify notations, we assume that preferences and costs are such that the north node will always export a net flow of electricity z_{ns} towards south. For this reason, $z_{ns} \geq 0$.

THE SOCIAL COST OF CONGESTION

Let η denote the dual variable associated to constraint (3). The first-order conditions that characterise the solution $(q_n^{go}, q_s^{go}, q_n^{wo}, q_s^{wo})$ to the above problem are:

$$U_s'\left(q_s^{wo}\right) = C_s'\left(q_s^{go}\right) \tag{4}$$

$$U_n'\left(q_n^{wo}\right) = C_n'\left(q_n^{go}\right) \tag{5}$$

$$\eta^o = C_s'\left(q_s^{go}\right) - C_n'\left(q_n^{go}\right) \tag{6}$$

$$q_n^{go} + q_s^{go} = q_n^{wo} + q_s^{wo} \tag{7}$$

$$\eta^o \geq 0, \quad K \geq q_n^{go} - q_n^{wo}, \quad \eta^o.\left[K - q_n^{go} + q_n^{wo}\right] = 0. \tag{8}$$

By (4) and (5), at each node marginal utility and marginal cost are to be equal. If the capacity of the line is very large, we obtain the "grid-free" allocation. In this case, $z_{ns}^o < K$ so that $\eta^o = 0$ by the complementary slackness condition (8). Then, by (6), the marginal valuation of electricity is the same at both nodes.

The series of equalities $U_s'\left(q_s^{wo}\right) = C_s'\left(q_s^{go}\right) = U_n'\left(q_n^{wo}\right) = C_n'\left(q_n^{go}\right)$, together with condition (7), fully describe the optimal allocation. (see Figure 4-4).

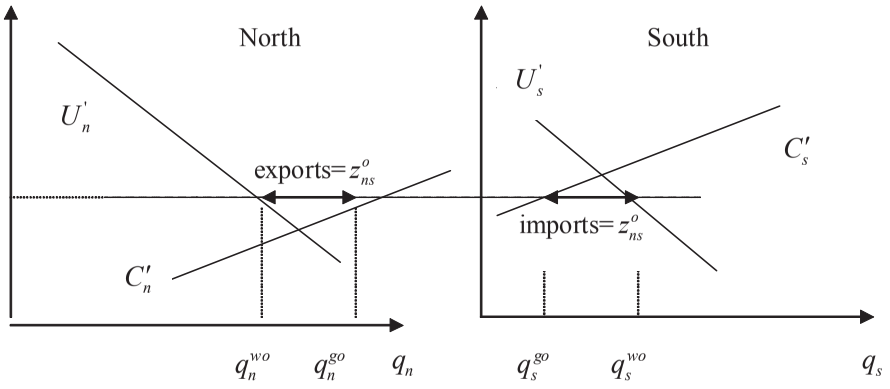

Figure 4-4. No capacity constraint, same valuation at both nodes

Conversely, if K is small, the constraint (3) is binding and we obtain the grid-constrained allocation. The optimal production in 'north' is given by $U_n'\left(q_n^{go}-K\right)=C_n'\left(q_n^{go}\right)$ and, in 'south', by $U_s'\left(q_s^{go}+K\right)=C_s'\left(q_s^{go}\right)$.

Now, the shadow price $\eta^o=U_s'\left(q_s^{go}+K\right)-U_n'\left(q_n^{go}-K\right)=C_s'\left(q_s^{go}\right)-C_n'\left(q_n^{go}\right)$ is strictly positive. It signals that a larger K would allow to generate more energy in 'north' where the cost of one additional unit is smaller than in 'south' and to consume more in 'south' where the utility of this additional unit is larger than in 'north' (see Figure 4-5).

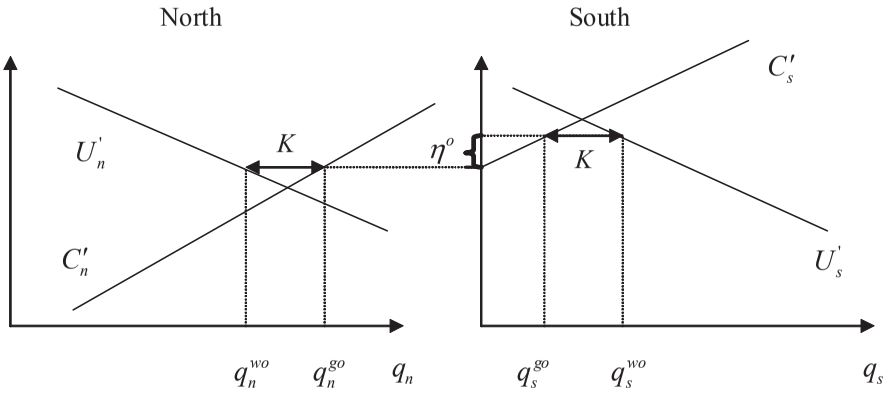

Figure 4-5. Binding transport capacity, nodal marginal valuations of energy diverge

THE SOCIAL COST OF LOSSES

Assume now that a fraction of the electricity injected at the north node is lost before arriving at the south node. In the program to obtain the first-best allocation, we must replace (2) with:

$$q_s^w-q_s^g=q_n^g-q_n^w-L\left(q_s^w-q_s^g\right) \tag{9}$$

since transmission losses $L(.)$ increase with the power consumed by the load. The first-order conditions remain (4), (5) and (8). But (7) is replaced by (9) with the consequence that (6) is replaced by:

$$\eta^o=C_s'\left(q_s^{go}\right)-C_n'\left(q_n^{go}\right)\left(1+L'\right) \tag{10}$$

where L' stands for marginal losses.

In the no-congestion case hereafter illustrated by Figure 4-6 ($\eta^o=0$), we see that the optimal allocation sets:

$$C_s'\left(q_s^{go}\right)=C_n'\left(q_n^{go}\right)(1+L')>C'\left(q_n^{go}\right).$$

This reflects that it takes (1+L') MWh generated in 'north' to match the need for 1 MWh in 'south'. As compared with the case where $L\equiv 0$, the volume of trade is reduced and

consumption in 'south' is smaller. But generation in 'north' can be larger in order to compensate for the losses.

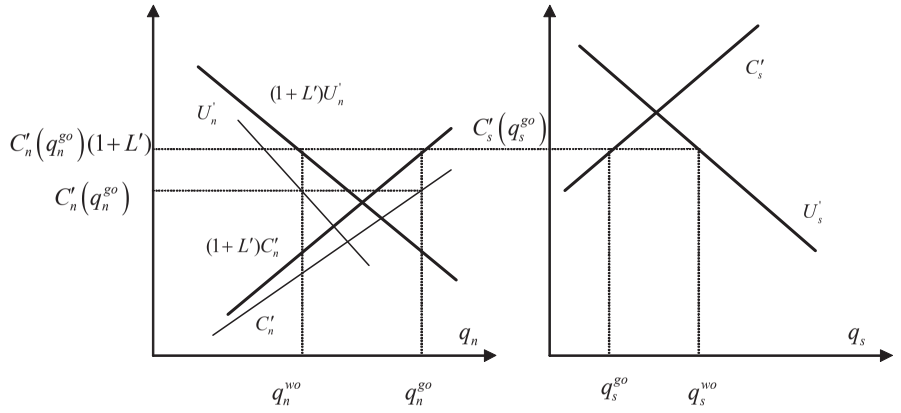

Figure 4-6. Optimal allocation with energy losses

TIME VARIATION

For a given transport infrastructure, the optimal dispatch is time dependent, following the changes in consumers preferences of and in generation plants availability. The obvious consequence is that nodal valuations of energy are also time-dependent. Assume there are no losses and the transport line of capacity K is used during two periods of equal duration: a peak period (labelled D for day) and an off-peak period (labelled N for night). The problem is to maximise:

$$\sum_{\tau=D,N} \sum_{i=n,s} [U_{i\tau}\left(q_{i\tau}^{w}\right) - C_{i\tau}\left(q_{i\tau}^{g}\right)] - rK \ ,$$

subject to two constraints of type (2) and two constraints of type (3), one pair for each period.[17] From the optimal allocation, we derive the shadow price of the line during the day $\eta_D^o = C'_{sD}\left(q_{sD}^{go}\right) - C'_{nD}\left(q_{nD}^{go}\right)$ and the shadow price of the line during the night $\eta_N^o = C'_{sN}\left(q_{sN}^{go}\right) - C'_{nN}\left(q_{nN}^{go}\right)$. As the need for transportation is higher during the day, we have $\eta_D^o \geq \eta_N^o$, where the equality holds only in the case $\eta_D^o = \eta_N^o = 0$, that is when the line is permanently non-binding (which should never occur since the unit cost of capacity is positive). Depending on the discrepancy between the night and day characteristics of consumption and generation, we can have either a permanently binding capacity ($\eta_D^o > \eta_N^o > 0$) or congestion only during the peak period ($\eta_D^o > \eta_N^o = 0$).

In any case, because the same line is used to satisfy different needs, it has the features of a "public good":[18] this explains why the test to know whether the capacity of the line is too large or too small consists now in comparing the unit cost of capital r with the sum ($\eta_D^o + \eta_N^o$).

IMPLEMENTATION

The first-best solution can be decentralised by means of nodal prices. At the south node any transaction will cost $p_s = U'_s\left(q_s^{wo}\right) = C'_s\left(q_s^{go}\right)$ per unit to a consumer and will return p_s

to a producer, whatever the location of the other side in the transaction. Similarly, at the north node $p_n = U_n'(q_n^{wo}) = C_n'(q_n^{go})$ will be charged for any unit of consumption and will be paid for any unit of production.

It is easy to check that consumers who solve:

$$\max_{q_i} U_i(q_i) - p_i q_i$$

will choose q_i^{wo}, and producers who solve:

$$\max_{q_i} p_i q_i - C_i(q_i)$$

will choose q_i^{go} at $i = n, s$. Therefore nodal prices, either resulting from competitive nodal markets or implemented by a regulation entity, allow to reach the first-best allocation.

When energy and transport are billed separately, the equilibrium price for transport from north to south must be such that $p_s = p_n + t_{ns}$; then $t_{ns} = p_s - p_n$. We conclude that, when losses can be neglected, from (6) the first-best transport price is:

$$t_{ns}^o = \eta^o \tag{11}$$

In words, absent any ohmic losses, transport should be free if and only if there is no congestion. Where there is congestion, a positive price is to be charged. Symmetrically, the equilibrium price for transport from south to north must be such that $p_n = p_s + t_{sn}$. We deduce that reverse flows should be charged $t_{sn}^o = p_n - p_s = -\eta^o \le 0$. Indeed, as long as $z_{ns} = q_n^g - q_n^w = q_s^w - q_s^g > 0$, individual transactions from south to north should be rewarded since they reduce the congestion of the line.

Similarly, taking losses into account, from (10) we obtain $t_{ns}^o = \eta^o + L'C_n' \ge 0$ and $t_{sn}^o = -(\eta^o + L'C_n') \le 0$. The latter is once more justified because any net injection in 'south' decreases the quantity of energy that must be injected at the north node and, consequently, decreases both the losses of energy and the congestion on the line.

BALANCING THE BUDGET OF THE TRANSPORT OPERATOR

Is this first-best sustainable on financial grounds? On the generator's side, if there are increasing returns to scale, marginal-cost pricing is not sufficient to break even. But since this report is dedicated to transport, we assume that generators do not lose money when generation is valued at first best[19].

What about the resources of the grid operator? The market equilibrium under nodal prices creates a "merchandising surplus":

$$\begin{aligned} MS &= p_s q_s^{wo} + p_n q_n^{wo} - p_s q_s^{go} - p_n q_n^{go} \\ &= \left[\eta^o + C_n' \cdot \left(L' - \frac{L}{K} \right) \right] K && \text{if the constraint is binding} \\ &= C_n' \cdot \left(L' - \frac{L}{z_{ns}^o} \right) z_{ns}^o && \text{otherwise.} \end{aligned}$$

This surplus is positive for two reasons:
because of the congestion rent $\eta^o K$ when the line is congested,
because losses are increasing with the square of the energy injected. It results that:

$$L' = 2\frac{L}{z^o_{ns}},$$

so that the surplus due to losses is $C'_n L$.

In summary, under nodal pricing the difference *MS* between the expenditures of customers and the gains of producers is positive and it can be used to cover the cost of the transport operator. But is it true that $MS \geq rK$?

If the size of the line could be optimally adjusted to transport needs, we would have $\eta^o = r$ and the *MS* would leave the operator with a net profit equal to the value of losses (provided there are no-increasing returns to scale). But for technical, economical and political reasons, in most transport networks, there is excess capacity so that $\eta^o < r$, and even $\eta^o = 0$ for absolute excess capacity. In most transport infrastructure, it results that the merchandising surplus does not cover the fixed costs.

We successively consider second-best linear prices and non linear prices.

RAMSEY NODAL PRICES

To simplify our analysis, we come back to the hypothesis of no losses. Then, the problem is to maximize (1) with respect to prices $(p^w_n, p^w_s, p^g_n, p^g_s)$ subject to (2) and (3), plus the additional obligation to balance the budget of the transport firm, that is:

$$p^w_n q^w_n + p^w_s q^w_s - p^g_n q^g_n - p^g_s q^g_s - rK \geq 0\,, \tag{12}$$

knowing that price-taking consumers behave optimally $\left(U'_i\left(q^w_i\right) = p^w_i\right)$ and price-taking generators behave optimally $\left(C'_i\left(q^g_i\right) = p^g_i\right)$ at each node *i=n,s*.

Denoting by λ the multiplier associated with the budget constraint (12), by γ the multiplier associated with the flow constraint (2) and by η the multiplier associated with the capacity constraint (3), from the first-order conditions, it is straightforward to write the second-best prices as:[20]

$$p^w_n - \frac{\lambda}{1+\lambda}\frac{p^w_n}{\varepsilon^w_n} = \frac{\gamma-\eta}{1+\lambda} = p^g_n + \frac{\lambda}{1+\lambda}\frac{p^g_n}{\varepsilon^g_n} \qquad \text{in 'north'}$$

$$p^w_s - \frac{\lambda}{1+\lambda}\frac{p^w_s}{\varepsilon^w_s} = \frac{\gamma}{1+\lambda} = p^g_s + \frac{\lambda}{1+\lambda}\frac{p^g_s}{\varepsilon^g_s} \qquad \text{in 'south'.}$$

Where:

$\varepsilon_i^w \overset{def}{=} -\dfrac{\Delta q_i^w / q_i^w}{\Delta p_i^w / p_i^w}$ is the price elasticity of demand and

$\varepsilon_i^g \overset{def}{=} \dfrac{\Delta q_i^g / q_i^g}{\Delta p_i^g / p_i^g}$ is the price elasticity of supply at node *i*.

Both consumers and generators at each node have to contribute to cover the fixed cost of transport. At node *i* we observe that:

$$p_i^w - p_i^g = \frac{\lambda}{1+\lambda}\left(\frac{p_i^w}{\varepsilon_i^w} + \frac{p_i^g}{\varepsilon_i^g}\right).$$

It is no longer true that there exists one unique price at each node. The difference between demand price and supply price increases with the value of the fixed cost of transport to be paid (which appears in $p_i^w - p_i^g$ through the dual variable λ) and it decreases when the demand and supply elasticities increase.

RAMSEY TRANSPORT PRICES

It obviously remains true that the equilibrium transportation charges applied to the dominant flow (from north to south) are such that $p_s^w = p_n^g + t_{ns}$. Using the Ramsey nodal prices, we obtain:

$$t_{ns} = p_s^w - p_n^g = \frac{\eta}{1+\lambda} + \frac{\lambda}{1+\lambda}\left(\frac{p_s^w}{\varepsilon_s^w} + \frac{p_n^g}{\varepsilon_n^g}\right) \tag{13}$$

As compared with (11), we observe two distortions:

— first, to the marginal cost of congestion[21] we have to add the contribution to the recovery of the transport fixed costs;

— second, the marginal cost of congestion is itself distorted.

An increase in *K* both increases λ for financial reasons and decreases η for technical reasons. It confirms that the pricing of transport mainly appears as the coverage of fixed costs, except when the capacity is very small.

Finally, note that for reverse transactions (from south to north) the Ramsey price shall be:

$$t_{sn} = p_n^w - p_s^g = -\frac{\eta}{1+\lambda} + \frac{\lambda}{1+\lambda}\left(\frac{p_n^w}{\varepsilon_n^w} + \frac{p_s^g}{\varepsilon_s^g}\right) \tag{14}$$

It means that the reward for alleviating losses and congestion will probably be more than compensated by the part of the tariff that participates to the coverage of fixed costs. If demand is much more elastic in 'south' than in 'north' and supply much more elastic in 'north' than in 'south', it can even be the case that $t_{sn} > t_{ns}$, despite the counter-flow effect. This is because financial concerns appear as dominant as compared with efficiency concerns.

In any case, comparing (13) and (14), we observe that $t_{ns} = t_{sn}$ only by coincidence. Actually, second-best prices are directional both for technical and financial reasons. Also, it

must be recalled that in (13) and (14) the dual variables, the nodal prices and the elasticities are all time dependent. We conclude that Ramsey transport prices must be time dependent.

TWO-PART TARIFFS

We consider now the case where prices are not constrained to be linear. We limit the analysis to the two-part tariff where $T_i^w(q_i^w) = a_i^w q_i^w + b_i^w$ is the expenditure for consuming q_i^w at node *i* and $T_i^g(q_i^g) = a_i^g q_i^g + b_i^g$ is the revenue when generating q_i^g at node *i*. The obligation to balance the budget of the transportation firm reads:

$$a_n^w q_n^w + a_s^w q_s^w - a_n^g q_n^g - a_s^g q_s^g + (b_n^w + b_s^w - b_n^g - b_s^g) - rK \geq 0 \tag{15}$$

By fixing $a_i^w = U_i'(q_i^{wo}) = a_i^g = C_i'(q_i^{go})$ at each node, we obviously implement the first-best allocation of consumption and generation at each node. Given these prices and given the flow constraint (2) (assuming *L*=0), (15) reads $(b_n^w + b_s^w - b_n^g - b_s^g) + (\eta^o - r)K \geq 0$. One obtains an infinite set of solutions for the fixed part of the locational energy tariffs, even when taking into account the participation constraints of consumers $U_i^w(q_i^w) - T_i^w(q_i^w) \geq \overline{u_i}$ and generators $T_i^g(q_i^g) - C_i^g(q_i^g) \geq \overline{g_i}$, where $\overline{u_i}$ (respectively $\overline{g_i}$) stands for the reservation value of consumers (respectively, generators) at node *i*. Any rule to share $(r - \eta^o)K$ among the agents by means of the fixed part of the tariffs that satisfies the above participation constraints, plus additional constraints due to the informational disadvantage of the price maker[22], is permitted.

When the buyer and the seller of a volume *q* are located at the same node *i*, they create a surplus:

$$T_i^w(q) - T_i^g(q) = a_i^w q + b_i^w - a_i^g q - b_i^g = b_i^w - b_i^g \tag{16}$$

To prevent any arbitrage between the two nodes, the trade of a quantity *q* of energy from north to south is to be billed $T_{ns}(q) = T_s^w(q) - T_n^g(q) = a_s^w q + b_s^w - a_n^g q - b_n^g$. Therefore:

$$T_{ns}(q) = \eta^o q + b_s^w - b_n^g \tag{17}$$

Symmetrically, an individual reverse transaction should be billed:

$$T_{sn}(q) = T_n^w(q) - T_s^g(q) = -\eta^o q + b_n^w - b_s^g \tag{18}$$

For the sake of transparency, one can prefer a uniform b^w for all consumers and a uniform b^g for all generators whatever their location, but the risk of exclusion (participation constraints that would not be satisfied) should not be neglected.

For the same reasons as the ones mentioned formerly, both the marginal price *a* and the fixed fee *b* should be time dependent. As long as participation constraints are not binding, the variation of *b* with the date is superfluous. In contrast, the variable part of the tariff for energy *a* and, consequently, the variable part of the tariff for transport η^o should be time dependent in order to send good scarcity signals to users.

MULTIPRODUCT TWO-PART TARIFFS

As already mentioned, energy prices and consequently transport prices should be varying with time. For example, the hourly two-part tariff defined in (17) should read:

$$T_{ns}(\tau, q_{ns}(\tau)) = A_{ns}(\tau) q_{ns}(\tau) + B_{ns}(\tau) \qquad (18)$$

and the total expenditures paid to the transport operator by an agent transferring energy from north to south, let us say during one year, would be:

$$T_{ns} = \sum_{\tau=1}^{8760} A_{ns}(\tau) q_{ns}(\tau) + B_{ns}(\tau) .$$

For the reason explained formerly, the fixed part of the tariff can be chosen quite easily[23], for example it can be a constant. The difficulty is that (18) requires to meter and to bill the flows from north to south almost continuously. Because the transaction costs would be very high, the operator will install simple metering and billing devices that do not distinguish the date of each flow. This means that the marginal price $A_{ns}(\tau)$ will not be a continuous function of time. In most cases, it will be a piecewise function with two seasonal (summer and winter) and two daily (day and night) values. At worst it will be uniform all the year long. Assume the latter.

In this case $A_{ns}(\tau) = A_{ns} \quad \forall \tau$ so that:

$$T_{ns}(\tau, q_{ns}(\tau)) = A_{ns} \sum_{\tau=1}^{8760} q_{ns}(\tau) + B_{ns}(\tau) = A_{ns} q_{ns}^{e} + B_{ns} ,$$

where $q_{ns}^{e} = \sum_{\tau=1}^{8760} q_{ns}(\tau)$ is the total energy consumed all the year long.

The obvious drawback of this two-part pricing is that it does not allow to discriminate between regular users who can be satisfied with a "medium size" grid on one side and, on the other side, irregular users who need large capacity of transport for short periods of time. A solution is to combine the transport demand for energy q_{ns}^{e} and a proxy for the transport demand for capacity, for example $q_{ns}^{K} = \max_{\tau} q_{ns}(\tau)$, and to bill them separately in a way that approximates the optimal time-dependent two-part tariff. We obtain a multi-product two-part tariff:

$$T_{ns} = A_{ns}^{e} q_{ns}^{e} + A_{ns}^{K} q_{ns}^{K} + B_{ns} \qquad (19)$$

LOOP FLOWS

Consider the three-node network of Box 4-2 and suppose that the line between the two injection nodes is the only one with a limiting capacity *K*. Neglecting losses, the optimal dispatch is the solution to:

$$\max_{q_1^g, q_2^g, q_3^w} \quad U\left(q_3^w\right) - C_1\left(q_1^g\right) - C_2\left(q_2^g\right) - rK$$

s.t. $\quad q_3^w = q_1^g + q_2^g$

$$\frac{\left|q_1^g - q_2^g\right|}{3} \leq K \, .$$

When the capacity constraint on this line is binding, each generator creates a counter-flow that alleviates the load provoked by the injection of power from the other plant. Because of this positive externality, the optimal dispatch can command that a high-cost plant should generate power despite the existence of available generation capacity at low cost somewhere else in the network. The level of the thermal constraint on the line between nodes 1 and 2 affects *all* nodal prices. If plant 1 is more efficient than plant 2, *i.e*, if $C_1'(q) < C_2'(q)$, prices:

$$p_2 \overset{def}{=} C_2'\left(q_2^g\right) = U'\left(q_3^w\right) + \frac{\eta^o}{3} > p_3 \overset{def}{=} U'\left(q_3^w\right) > p_1 \overset{def}{=} C_1'\left(q_1^g\right) = U'\left(q_3^w\right) - \frac{\eta^o}{3}$$

allow to encourage (discourage) generation at node *2* (node *1*).

If energy and transport have to be invoiced separately, from the equilibrium conditions $p_j = p_i + t_{ij}$, we obtain the set of linear tariffs for transport:

$$t_{13} = \eta^o / 3 \, , \qquad t_{23} = -\eta^o / 3 \, , \qquad t_{12} = 2\eta^o / 3 \, .$$

They explicitly encourage consumers to transact with expensive generators (node 2) and they dissuade them to transact with low-cost generators (node 1). These prices send the accurate signals to prevent jeopardising the safety of the grid, namely the ohmic constraint on the line between nodes 1 and 2.

Note that, in contrast to the one-line model, in a meshed network the shadow cost of the thermal constraint on the line between two nodes is no longer equal to the difference in marginal valuations at the nodes. In our illustration, $\eta^o = 3/2\left(C_2'\left(q_2^g\right) - C_1'\left(q_1^g\right)\right)$. This is because all connected lines are affected by a bilateral transaction. Consequently, the marginal value of congestion reflects all the negative externalities created through the grid by the limited capacity of each piece of equipment.

REFERENCES

Chao, H.P. and S.Peck (1996), "A Market Mechanism for Electric Power Transmission", *Journal of Regulatory Economics*, vol 10, pp 25-60.

Chao, H.P., S.Peck, S.Oren and R.Wilson (2000), "Flow-Based Transmission Rights and Congestion Management", *Electricity Journal,* October, pp 38-58.

Crampes C. and J.J.Laffont (2001) "Transport Pricing in the Electricity Industry" *Oxford Review of Economic Policy*, Autumn, vol 17, no 3, pp 313-328.

Cremer, H. and J.J.Laffont (2002), "Competition in Gas Markets", *European Economic Review*, vol 46, no 4-5, pp 928-935.

Hogan, W.W. (1992) "Contract Networks for Electricity Power Transport", *Journal of Regulatory Economics* vol 4, pp 211-242.

Hogan, W.W. (1998) "Nodes and Zones in Electricity Markets: Seeking Simplified Congestion Pricing", in Hung-Po Chao and H.G. Huntington (Eds.) *Designing Competitive Electricity Markets,* Kluwer Academic Publishers, London, pp 33-62.

Hsu, M. (1997) "An Introduction to the Pricing of Electric Power Transmission", *Utilities Policy*, vol 6, no 3, pp 257-270.

Joskow, P.L. and J.Tirole (2000) "Transmission Rights and Market Power on Electric Power Networks", *RAND Journal of Economics*, vol 31, no 3, Autumn, pp 450-487.

Pérez-Arriaga I.J., F.J.Rubio, J.F.Puerta, J.Arceluz and J.Marín (1995) "Marginal pricing of transmission services: an analysis of cost recovery", *IEEE Transactions on Power Systems,* vol 10, no 1, February, pp 546-553.

Schweppe F., M.Caramanis, R.Tabors and R.Bohn (1988) *Spot pricing for electricty*, Kluwer Academic Publishers.

Stoft S. (2002) *Power System Economics*, IEEE/Wiley.

Wu, F., P.Varaiya, P.Spiller and S.Oren (1996), "Folk Theorems on Transmission Open Access: Proofs and Counter examples", *Journal of Regulatory Economics*, pp 5-23.

NOTES

[1] This is a good illustration of the difficulty to give a precise, objective definition of costs, as explained in Chapter 2.

[2] As shown in Chao and Peck (1996), physical rights and financial rights are equivalent when the energy markets and the rights markets are perfectly competitive. Joskow and Tirole (2000) provide an analysis of various non-competitive configurations.

[3] See Joskow and Tirole (2000), p. 452.

[4] For an illustration see Box 4-2. For more details, see for example Hsu (1997).

[5] Schweppe *et al.* (1988).

[6] See for example Stoft (2002), p. 417.

[7] Studies of market power are based on calibrated simulation models. See references in Crampes and Laffont (2001).

[8] We do not consider the issue of nodal-price randomness. To hedge against price volatility, users can sign financial contracts. See Hogan (1992) and Chapter 5 *infra*.

[9] See Oren *et al.* (1996).

[10] See Pérez Arriaga *et al.* (1995).

[11] Provided the authority in charge of pricing is not constrained by a lack of information on the willingness to pay of the users. For an illustration, see Chapter 6.

[12] See Chapter 3.

[13] For details on economic and legal price discriminations, the reader is referred to Chapter 1.

[14] On the principles of cooperative-games theory, see Chapter 3.

[15] In the two-part case, the variable part is increasing linearly with the quantity of energy transported. In multi-part tariffs, the variable part is a piecewise linear increasing function.

[16] Details on actual tariffs are presented in Chapter 8.

[17] We do not discuss the additional intertemporal constraints due to the management of hydroplants or to the obligation to satisfy heating conditions in thermal plants.

[18] Actually, D and N are successive, but the argument is the same as if there were several non-rival simultaneous needs to satisfy with the same equipment: it is the willingness to pay of *all* users that must be taken into consideration.

[19] This is true when the long run marginal cost of generation is non decreasing and generators have optimized the size of their plants.

[20] See Cremer and Laffont 2001.

[21] To which the marginal cost of losses should be added.

[22] Self-selection constraints are to be added to design the menu of tariffs when the price maker cannot observe some individual characteristics of the price takers, for example the consumers' willingness to pay or the generation cost of producers. The problem is to prevent an opportunistic switch of some users towards tariffs that are designed for somebody else. Chapter 6 provides an illustration of the design of control mechanisms under information asymmetry.

[23] Except when there is a serious informational gap that imposes fine tuning to respect all the participation and self selection constraints.