Transmission pricing and performance-based regulation

Ingo Vogelsang, Boston University*

Paper Prepared for CarnegieMellon Conference on Electricity Transmission in Deregulated Markets: Challenges, Opportunities, and Necessary R&D Agenda Pittsburgh, December 15-16, 2004

Abstract

Performance-based regulation (PBR) is influenced by the Bayesian and non-Bayesian incentive mechanisms. While Bayesian incentives are impractical, the insights from their properties can be combined with practical non-Bayesian mechanisms for application to transmission pricing. This combination suggests an approach based on the distinction between ultra-short, short and long periods. For ultra-short periods real-time pricing of point-to-point transmission services with the help of an ISO would be adequate. Pricing in short periods involves the calculation and payment of fixed fees and adjustments via RPI-X type formulas or profit sharing. While short-term pricing occurs under full commitment by the regulator, productivity-enhancing incentives may have to be tempered by long-term consideration, so that profit sharing may be preferable to pure price caps. Long periods mark the limits of regulatory commitment and are still short relative to network investments. As a result, incentives should be further weakened by adjustments based on rate-of-return regulation with a "used and useful" criterion.

* The author's research on this project was supported by NSF Grant ECS-0323620 to Carnegie-Mellon University on "EPNES: Dynamic Transmission Provision and Pricing for Electric Power Systems."

1. Introduction

Performance-based regulation (PBR or, rather, incentive regulation) is characterized by two main properties. First, it gives the regulated firm some behavioral discretion, for example, in the choice of prices. Second, it rewards (punishes) the firm for improving (deteriorating) performance relative to the regulator's objectives. In addition, it is often more compatible with the opening of regulated markets to competition than traditional cost-of-service (or rate-of-return) regulation. The current literature on PBR developed from two very different needs.

The first need was to improve the perceived theoretical and practical inadequacies of U.S. rate-of-return regulation. This part of the literature had its roots in the discovery of the Averch-Johnson effect, according to which rate-of-return regulation deviates from cost minimization (Averch and Johnson, 1962). But it began in earnest only with Baumol's "plausible policies for an imperfect world" (Baumol, 1967), leading to new developments such as price caps and yardstick regulation. This literature had a substantial impact on the regulation of network industries worldwide and particularly on telecommunications. PBR is also common in U.S. telecommunications but not as much in electricity. This lack of PBR for the electricity sector may be due to lumpy and long-term investments, which are hard to handle with incentive schemes of shorter duration, or due to the prevalence of cost increases (rather than declining costs) over time, which makes incentive regulation look unattractive.

The second need came from the theoretical literature on optimal pricing for public enterprises and regulated industries that had been developed by Hotelling (1938) and Boiteux (1956). This literature had come up with marginal cost pricing and Ramsey pricing rules but had neglected the lack of information of regulators to implement such rules in light of the informational superiority of the regulated firms. During the 1970s the principal-agent framework had been developed to deal with similar problems of asymmetric information in the context of managerial incentives under separation of ownership and control. The merger of the principal-agent approach and the optimal pricing literature then lead to the Bayesian approach to incentive regulation by Baron and Myerson (1982), Sappington (1983) and Laffont and Tirole (1986). This literature has grown substantially over time but has had very little concrete and visible impact on the way regulation has been done. One reason for this is the difficulty to translate the approach into rules that regulators can apply directly. Another reason may be that some of the potential learning from this approach is not viable in the actual regulatory environment (Crew and Kleindorfer, 2002).

While this short characterization of the two approaches would suggest that transmission pricing be best served by the practical approach of plausible rules, we will argue below that both approaches are needed to deal with transmission pricing problems.

Transmission pricing is a complicated task. The lack of storability of electricity, in combination with transmission capacity constraints, suggests that pricing in the very short run is important in order to avoid congestion. At the same time prices have to guide

operating and investment decisions by transmission companies, generators and load-serving entities. This suggests pricing approaches geared at highly differentiated time horizons. Furthermore, transmission cost functions are affected by loop flow problems, power losses and ancillary services.

2. Bayesian vs. non-Bayesian incentive schemes

2.1 Bayesian incentive schemes

2.1.1 Characterization

Under the Bayesian incentive approach the regulator is viewed as a principal who uses the regulated firm as an agent in order to fulfill the principal's objective, which is well-defined and expressed in monetary terms. A typical regulatory objective would be to maximize $W(p) = V(p) + \alpha\pi(p)$, where $W(p)$ represents welfare as a function of the firm's price, $V(p)$ is consumer surplus, α is a weight between 0 and 1, and $\pi(p)$ is the firm's profit. The tool is usually a monetary transfer $T(p)$ from the regulator with which the firm is induced to act in the principal's interest. Since the transfer occurs between the regulator and the firm, it affects the regulator's objective function not only through the behavioral effect on the firm but also directly. Here the assumption is that the effect on the firm's profit is weighted by α, while the effect on the regulator's budget is weighted by 1.

Information is asymmetric in that the agent has important knowledge about herself or about the situation that the principal does not have. Items that matter for performance, such as the firm's effort to reduce costs, are framed in terms of asymmetric information. If the regulator could costlessly observe such effort he would hold the firm to the optimal effort level directly. This asymmetry of information is captured by the assumption that the principal has probabilistic information about the agent (knows the distribution of types that the agent belongs to), while the agent knows her actual type. The principal and agent are further assumed to be fully informed about all the other relevant properties (functional forms, etc.) necessary to fulfill the principal's objective.

Since the agent acts in her own interest by maximizing $\Pi(T(p), p) = \pi(p) + T(p)$, principal and agent play a game against each other. One specific feature of the principal-agent game is that the principal usually moves first by setting the incentive scheme for the agent. In a single-period setting the principal therefore solves the game backwards by taking into consideration the agent's response to the principal's incentive scheme. This response is captured by two constraints on the principal's optimization problem. The first is that the agent will only be willing to act on the principal's behalf if she is rewarded at least with her reservation utility (participation constraint = PC), which is usually normalized to zero.[1]

[1] Since the principal is maximizing his expected objective over the distribution of the agents types, it is not clear that the participation contraint is required in its usual form. It makes the principal totally risk averse against the possibility of not attracting any agent. Such total risk aversion may be a reasonable assumption.

The second constraint is that the agent is going to maximize her own utility under the incentives provided by the principal. Since these incentives depend on the actual type of the agent they have to be designed in such a way that the agent is induced not to mimic someone else's type (incentive compatibility constraint = ICC).

In the case where the regulated firm's type (θ) and effort (e) cannot be observed and where the unobservable cost of effort to the firm is ψ(e) the maximization problem for the regulator is to find an incentive scheme T(p) with

$$\begin{aligned} T(p) &= \arg\max \int_{\theta} [V(p(\theta) + \alpha\pi(p(\theta)) + (\alpha - 1)T(p(\theta))] f(\theta) d\theta \\ s.t. \quad PC: \quad \Pi(p(\theta), T(p(\theta)) &\geq 0 \quad and \quad ICC: \quad \partial\Pi / \partial\theta = -\psi'(e(\theta)) \end{aligned} \tag{1}$$

The ICC results from the observation that a more efficient type θ^+ can mimic a less efficient type θ^- simply by exerting less effort and still making the same profit and getting the same transfer. As a result, mimicking type θ^- would give type θ^+ an extra profit from the reduced cost of effort.

The incentive schemes designed under this approach are called Bayesian because the probability distribution of types used by the regulator is subjective with density f(θ) and cumulative distribution function F(θ) and is updated (in a multi-period problem) according to Bayesian rules of probability updating. The prior probabilities are simply assumed as given.

Crew and Kleindorfer (2002) severely criticize (a) that the approach spends no intellectual effort on the derivation of the probability distribution F(θ) and (b) that F(θ) is assumed to be common knowledge. Both these criticisms are severe. However, some strands of the Bayesian literature deal with regulatory monitoring that can be interpreted as the gathering of information about agents' types and thereby dealing with the first criticism (Baron and Besanko, 1984). Nevertheless, Crew and Kleindorfer are right that no part of the Bayesian literature reflects the actual information-gathering process of regulatory procedures. The second criticism may be countered by the argument that the regulated firm does not have to have the same information about the distribution as the regulator as long as she knows her own type.[2] However, even if no common knowledge assumption is required the assumption of subjective priors means that the scheme cannot be monitored by the general public that employs the regulator (Vogelsang, 1988). Crew

---

However, it could well be that the principal prefers the chance of higher realization of his objectives for more efficient types of agents going along with a small risk of zero fulfillment of his objective for the less efficient types over some objective fulfillment for all types. As already shown by Baron and Myerson (1982), the results of this alternative approach would not change much of the current literature provided the difference between fulfillment of the principal's objective by the least efficient agents and excluding these agents altogether is sufficiently large.

[2] The Bayesian approach normally assumes that the regulator has full information about most of the firm's problem, e.g., about the functional forms of demands and costs.

and Kleindorfer are therefore right that the use of the principal-agent framework is suspect because the regulator lacks the sovereignty of a real principal. The answer already anticipated by the Bayesian incentive literature has been to complicate the model by treating regulators simultaneously as principals and agents in a two-stage principal-agent framework, where regulators are employed by the general public and themselves employ the regulated firm (Laffont and Tirole, 1993, chapters 11 and 15).

A further criticism of the Bayesian approach is its use of transfers to the firm as the main tool for influencing the regulated firm's behavior.[3] Normally, regulators do not have the authority to pay transfers to firms. Nor can they tax firms if the transfers turn out to be negative (which they often are in the optimum).[4] The transfer issue, however, can often be solved by making consumers pay fixed fees equal to these transfers. As long as the fixed fees do not exclude consumers, their allocative effects are negligible (depending on income effects or the like). But if they are charged in conjunction with and in addition to other fixed fees that is not always a safe assumption.

The more realism is included in the regulatory problem under the Bayesian approach the less feasible it becomes to derive directly applicable quantitative regulatory rules. Rather, the results of the Bayesian approach only lend themselves to general qualitative prescriptions that have to be filled out by the legislature and by concrete regulators. In my view, this is best done in combination with the simple rules derived under the non-Bayesian approach. The results from the Bayesian literature could determine the design of non-Bayesian mechanisms and the choice among several non-Bayesian mechanisms.

2.1.2 Main results and insights

What are the qualitative insights of the Bayesian approach?

(1) The full information optimum is generally not achievable. The claimed exception is the absence of distributional concerns ($\alpha = 1$) and of any costs of raising public funds for transfers. The example for that is the Loeb-Magat mechanism (Loeb and Magat, 1979), which, however, additionally requires full information of the regulator about the demand curve. Regulation is always imperfect. So is incentive regulation. The question is therefore primarily if it is an improvement compared to traditional cost-of-service regulation.

(2) The quality of Bayesian regulation depends crucially on the ability and intentions of the actual regulator. This requires that the application the regulation schemes be based on mutually observable data, meaning that incentives have to be compatible with regulatory accounting. The better the hard information of the regulator the better the goal achievement and the lower will generally the information rents be that have to be granted to the firm.

[3] This is not a problem in the other application of this literature which is for government procurement.
[4] For example, under the original Baron-Myerson mechanism the least efficient firm type charges a price above costs but is just breaking even, due to a transfer to the regulator (Baron and Myerson, 1982).

(3) The most important insight is that superior efficiency by the firm (for revealing information and for superior effort) should be rewarded by increased profits. The profit or information rent should increase in the innate efficiency of the regulated firm (its type). This is the direct result of the PC and the ICC. The participation constraint requires that any type of firm earn nonnegative profits. Because any more efficient firm type can always mimic the most inefficient type, the incentive compatibility constraint then assures that any but the most inefficient type will earn positive profits. This also means that, from the regulator's perspective, expected profits (with expectations over the θ's) are positive. Crew and Kleindorfer (2002) severely criticize this feature of Bayesian schemes, claiming that such positive profits would be politically unrealistic. Also, Joskow and Schmalensee (1986) hold that the firm should be able to make positive profits under some circumstances, but negative profits under others, with at least cost coverage over time. While this recommendation appears to be contrary to most of the Bayesian incentive literature, it hinges on the interpretation of the participation constraint.[5] It may be that zero economic profits do not represent the reservation utility of the firm. The reservation profit could, for example, be negative after entry and in the presence of sunk costs. The longer the time horizon, the closer the reservation utility will be to zero profits. It also is worth noting that almost all of the literature on the Averch-Johnson effect assumes that allowed profits under rate-of-return regulation exceed the cost of capital to the firm. This claim is theoretically backed by Evans and Garber (1988), while it is called into question by Joskow and MacAvoy (1975).

(4) The higher the perceived innate efficiency of the firm (the more efficient a firm's type) the more cost-reducing effort should it be induced to exert. This again follows from the ICC, because the more efficient types can mimic the less efficient types by exerting less effort. Since excessive profits are disliked by the regulator (rent extraction), the regulator is only willing to trade off such profits against productive efficiency. Combined with the incentive compatibility requirement this means that the firm is best offered a menu of regulatory options that lead to a self selection among firm types and accompanying effort levels.[6] The menu would be such that the lower cost types would choose the higher-powered incentive (for example, a price cap), while the higher cost types would choose the lower-powered incentives (for example, profit sharing or rate-of-return regulation).

While the FCC has tried out price regulation with menus for the Bell operating companies in the early 1990s, this practice was abandoned after a few years. I am not aware of any serious discussion of the menu approach in practice and will therefore not discuss it further here.

[5] Because of the exercise of discretion under PBR the regulated firm's risk is usually increased. This can lead to losses from time to time and those would have to be compensated by the chance of making higher profits. Further the firm's cost of capital may increase as a result of PBR, lowering expected economic profits.

[6] The distribution of types may actually be altered by the introduction of incentive schemes because more incentive-oriented managers would be attracted to regulated firms.

(5) The Bayesian literature provides few insights about pricing. The strongest appears to be the incentive-pricing dichotomy (Laffont and Tirole, 1993), according to which (under some plausible conditions) pricing can be done optimally independent of the incentives for cost reduction. This result, however, is only valuable if the regulator can implement optimal prices based on revealed costs.

There are no sharp results on pricing with unknown demand. When the optimal full information price is unequal marginal costs no general policy under asymmetric information about demand (and costs) can be formulated (Armstrong and Sappington, 2003). This lack of direction provided by this literature would be relevant under economies of scale and if no Coase tariff were feasible.

The results on pricing suggest that the non-Bayesian approaches receive little guidance from the Bayesian literature on pricing for allocative efficiency except that there exist no general reasons against aiming for allocatively optimal prices.

(6) The ability of incentive regulation to achieve the regulatory objective depends on the number of instruments available relative to the number of objectives. Thus, price regulation combined with a tax/subsidy scheme can better achieve allocative and productive efficiency that pricing alone. Two-part tariffs can achieve more than linear tariffs, in particular if the fixed fee has no allocative effects on consumers.

Because quality improvements are costly, incentives for cost reductions and quality concerns can be in conflict. If price setting is the only policy instrument for both objectives cost-cutting incentives may have to be reduced if quality concerns are high. Preferable, however, is the application of specific quality-oriented instruments. Because we consider price regulation only, we will therefore leave out quality aspects from explicit consideration, assuming that quality optimization will be pursued with a separate instrument.

(7) Commitment of the regulator emerges as an important driver of the power or steepness of incentives provided optimally. The power or steepness of incentives is the degree to which the firm's revenues under the mechanism are independent of costs. The less revenues depend on cost the steeper or more high-powered the mechanism. Credibility of commitment has to be traded off against steepness of incentives. Commitment mostly refers to future periods. Thus, if commitment is restricted to the first period then this only allows for steep incentives if the activity (and/or the relevant firm type) only lasts for that period. If, however, the activity or validity of type information extends beyond that period less steep incentives are optimal. Joskow and Schmalensee (1986) therefore recommend against extending incentive regulation to long-run investment decisions. The less regulatory commitment, the flatter are the optimal incentives. As a result, rate-of-return regulation does not look so bad after all. It may just be an imperfect way to achieve optimality for long-run investments under a variety of realistic constraints.

Further specific findings on commitment are:

(7a) Under full commitment the regulator does not use its learning to update the policy within the time of commitment.
(7b) Renegotiation (commitment to the firm to maintain its rent) is generally not optimal.
(7c) Under an inability to commit, less information by the regulator may be better than more information. It is hard to see how such an advice can be purposely implemented.
(7d) Commitment is also bad if regulators are myopic, because they then may impose costs on society in the future (Lewis and Sappington, 1990). For example, the regulator could provide short-term incentives to reduce costs by deferring maintenance and thereby increase costs later.

(8) The results in (7) suggest that increasing regulatory commitment would be desirable. However, restrictions on regulatory commitment are often deliberate in order to hold regulators accountable and reduce regulatory capture. Regulatory capture leads to a divergence between the regulator's and society's objective: Under capture regulatory commitment may therefore be bad, because then the result of capture cannot be undone. The danger of capture can therefore be one of the legitimate reasons for limits on regulatory commitment. It also explains restrictions on the regulatory tools, for example, the inability to use taxes and subsidies.

(9) Outcomes should be tied to variables that are under managerial control as opposed to variation of outside variables (unless the two cannot be separated and regulators are risk neutral).

(10) Last, the results of the Bayesian literature vary substantially by the assumed circumstances. "The fact that information, technology, instruments, and institutions all matter in the design of regulatory policy implies that the best regulatory policy typically will vary across industries, across countries, and over time." (Armstrong and Sappington, 2003)

2.2 Non-Bayesian incentive schemes

In our discussion of non-Bayesian incentive regulation we leave out a number of suggested and practiced approaches, such as franchise bidding, yardstick regulation and double sourcing. We also do not consider substitutes for transmission in the form of distributed generation. Reliability and other quality aspects are assumed given.

Non-Bayesian approaches are based on the assumption that the regulated firm has superior information (about costs and/or demands) to the regulator and that this superiority can be converted into additional surplus by giving the firms behavioral discretion to use this information in a way that improves performance under the regulator's objectives. This presupposes that the mechanism achieves some alignment of the firm's objective (in profits) and the regulator's objectives. The approaches are therefore linked to variables that are closely related to social surplus, the perceived regulatory objective function of this literature. Accordingly, firms can receive either the whole social surplus (Loeb-Magat, 1979), a transitional gain of the whole surplus

increase (Sappington and Sibley, 1988) or of part of the surplus increase (Vogelsang and Finsinger, 1979).

Under the Loeb-Magat scheme (1979) the firm receives the consumer surplus as a subsidy and therefore maximizes the regulatory objective directly (provided there are no distributional weights). Since distributional weights have to be assumed for reasons of fairness and since subsidies have to be ruled out and since consumer surplus cannot be observed with any degree of accuracy, other mechanisms have played a bigger role as suggestions for actual applications.

The closest to the Loeb-Magat scheme is the incremental surplus subsidy scheme (ISS) by Sappington and Sibley (1988). The ISS belongs to the class of dynamic schemes with more direct potential application. According to the ISS the firm would receive a subsidy in each period (or pay a tax) equal to the difference between the consumer surplus increase in the current period and last period's profit: $ISS^t = V(p^t) - V(p^{t-1}) - \pi(p^{t-1})$.[7] Sappington and Sibley show that this mechanism converges to welfare-optimal prices (assuming equal weights) in a single period. Thus, the outcome is virtually always efficient. Three potential drawbacks are (1) that consumer surplus change needs to be observed by the regulator, (2) that the firm may earn rents in the first period and (3) that the regulator needs to be able to subsidize or tax the firm. The last of these drawbacks could be eliminated by the introduction of a fixed fee that lets consumers rather than the regulator "subsidize" the firm.[8] Provided no consumers exit the market because of the fixed fee, the incentive properties would be maintained. The second property is likely to be part of any incentive mechanism and therefore only of special concern if the rent becomes large. The first drawback could be overcome by observable approximations to the consumer surplus change. The simplest such approximation (the Slutsky approximation) is used in Finsinger and Vogelsang (1982), which is otherwise similar to the ISS. In this case the subsidy would be $S_{FV}^t = (p^{t-1} - p^t)q^{t-1} - \pi(p^{t-1})$. This process leads to a convergence to welfare-optimal pricing in a stationary environment in the long run. Because the environment is likely to change over time, this delayed convergence is an undesirable property which could worsen through strategic behavior by the firm. A closer approximation to the consumer surplus change would therefore be desirable. Vogelsang (1988) shows that closer approximations are likely to be subject to strategic manipulations (cyclical behavior) by the firm. However, this disturbing outcome only holds if demand is nonlinear and if investments can be undone. Since transmission investments are largely sunk and since demands are often linear in the relevant price range, a simple observable second order approximation to the consumer surplus change may be the way to go.

The ISS, based on two-part tariffs, could be interpreted in the tradition of the price-cap approach. Under price caps there are no subsidies and the regulated firm is allowed to charge prices only that obey a price-cap formula. This formula is usually supposed to

[7] Sappington and Sibley calculate profits based on expenses rather than costs. As a result, the regulator may be asked to finance investments.

[8] Negative fixed fees (= taxing the firm) are problematic, though, because they could induce customers to split up their purchases to collect negative fixed fees several times over.

provide incentives for cost minimization, investments and allocative efficiency, while keeping the firm's rents at reasonable levels. The first mechanism trying to achieve such a combination of objectives was the RPI-X scheme by Littlechild (1983) that combined properties on cost reduction originally developed by Baumol (1982) with properties on allocative efficiency developed by Vogelsang and Finsinger (1979):

$$p^t q^w / p^{t-1} q^w \leq (1 + i - X) \qquad (2).$$

Here, p and q are price and quantity vectors of outputs, and superscripts t refer to time period and w to weights. The left-hand side of (2) is a price index. Chained Laspeyres weights ($q^w = q^{t-1}$) are most common in price cap applications, because they are easily calculated and have fairly good economic properties. Idealized weights ($q^w = q^*$) with strong efficiency properties would equal perfectly predicted quantities (Laffont and Tirole, 1996). The right hand side of (2) determines the price level. 'i' is the rate of inflation[9] and 'X' is the adjustment factor set by the regulator. In a stationary environment the price-cap process with chained Laspeyres weights will converge to a Ramsey price structure. However, profits will generally be positive at the point of convergence. Under idealized weights the Ramsey price structure will be reached in a single period. However, the regulator would have to find the weights correctly.

The price-cap approach expressed by equation (2) with Laspeyres weights is actually quite related in spirit to the ISS. (2) can be reformulated as

$$(p^{t-1} - p^t)q^{t-1} - p^{t-1}q^{t-1}(X - i) \geq 0 \qquad (2').$$

The first term on the left hand side is the Slutsky approximation of the change in consumer surplus, while the second term can be interpreted as a correction for excess profits (if $X > i$) or insufficient profits ($X < i$). Now the firm would maximize the Lagrangian

$$\text{Max } L = \pi(p^t) + \lambda\{(p^{t-1} - p^t)q^{t-1} - p^{t-1}q^{t-1}(X - i)\} \qquad (3).$$

If the constraint is binding λ is positive, meaning that the firm would be maximizing an expression that is related to the change in social surplus. It would be a direct approximation of the change in social surplus if $\lambda = 1$ and if $p^{t-1}q^{t-1}(X - i)$ could be interpreted as excess profit in the past period.

An alternative to price caps is profit sharing that would, at least in the short run, provide softer cost-reducing incentives than price caps. Under profit sharing instead of price caps X would be replaced in (2) by $s\pi^{t-1}/p^{t-1}q^{t-1}$, where π is profit and s is the profit share going to consumers.[10] The resulting process would, in a stationary environment, converge to

[9] This would correspond to the change in the retail price index (RPI) in the UK, from which the RPI-X formula got its name.

[10] This becomes the mechanism in Vogelsang and Finsinger (1979) for s = 1 and i = 0. On a further discussion of profit sharing (sliding scale), see Navarro (1996), p. 114 and 136, 138. Navarro in particular discusses progressive vs. regressive sharing.

Ramsey prices under chained Laspeyres weights. Here, profits would also converge to zero. The reason for the zero profits is that prices are changed from period to period, as long as profits are nonzero (positive or negative). Thus, convergence presupposes zero profits.

This short characterization of non-Bayesian incentive schemes suggests that one can separate revenue requirement (= price level) from rate design (= price structure). The revenue requirement is expressed in the starting value of a price index and in an (RPI – X)-type or profit-sharing adjustment. The rate design is expressed in the chosen index weights and possible constraints on price restructuring.

There are benefits and risks of granting pricing flexibility to the regulated firm. Some flexibility is optimal because of the superior information of firm on costs and demands. The type of flexibility depends on the type of index used to constrain price level. Examples include average revenue constraints, chained Laspeyres weights, idealized weights, combined Laspeyres/Paasche weights etc. The flexibility of rate design can also extend to the introduction of two-part tariffs and more complicated non-linear tariffs. Using a two-part tariff, Vogelsang (2001) shows that setting pricing flexibility can mean flexibility of investment (due to ISO pricing). Tradeoffs can exist between deviations from marginal-cost pricing and productive incentives, because the regulated firm may have to be able to reap the benefits from cost reduction in the form of prices that exceed costs (Schmalensee, 1979). In Vogelsang (2001), however, all the productivity incentives are captured in the fixed fees. Provided the size of the fixed fee does not affect the number of transmission users, allocative efficiency can be achieved over time.

Generally a tradeoff exists for non-Bayesian mechanisms between implementability (verifiability at reasonable costs = fulfillment of informational and transactional constraints) and achievement of the regulatory objectives (and fulfillment of administrative constraints on regulators, e.g., no subsidies or taxes and no commitment beyond a certain time horizon).

The non-Bayesian mechanisms tend to be dynamic adjustment processes with some optimality properties in the limit, as time goes to infinity, but otherwise only improvements over the status quo. However, because of changing environments, revisions of the processes are required from time to time. These revisions usually coincide with the time limits, beyond which regulators cannot commit anyhow. This leads to the question of guidance from the non-Bayesian literature on these revisions. Beesley and Littlechild (1989) and Crew and Kleindorfer (1996) discuss the practical approaches to this problem in the UK and U.S. However, there do not exist genuine non-Bayesian regulatory mechanisms for the long run. The only applicable work in the theoretical literature seems to be Gilbert and Newbery (1994). They show that rate-of-return regulation with a "used and useful criterion" can sustain optimal investment. A "used and useful" criterion with prudence reviews improves investment incentives by avoiding over-capitalization of the Averch-Johnson type. While the non-Bayesian incentive mechanisms are designed for allocative and productive efficiency improvements in the short run, there does not therefore appear to exist a mechanism for

long-term improvement over the rate-of-return regulation alternative (with the "used and useful" proviso). The results of the Bayesian literature on regulatory commitment could provide an explanation for this lack. All that non-Bayesian mechanisms can hope to do in this respect is to lengthen the commitment period.

3. Application to transmission pricing

3.1 Distinguishing features of transmission markets and their regulation

Transmission services are characterized by a number of peculiarities that affect the appropriate type of regulation. The peculiarities include very distinct timing issues, a superiority of monopoly supply and vertical economies with generation and load-serving entities (LSEs).

Foremost among and at one extreme of the timing issues is the requirement for real time coordination between generation and loads due to the equilibrium requirements of transmission networks. At the other extreme of the time scale are long-term investments in network facilities, particularly in high-voltage lines, that are sunk and possibly lumpy.

A particularly delicate problem comes from the fact that capacity costs are a Transco's paramount cost factor, while the services are largely nonstorable. As a result, the Transco faces the problem of optimal capacity expansion and capacity utilization. While expansion requires cost coverage and stable signals, utilization requires fluctuating prices that follow demands. Prices that preserve long-run investment incentives can, for example, be oriented at long-run cost estimates. The long-term investments exhibit economies of scale and scope, but loop-flow problems can also cause potential diseconomies of scope. Transmission has complicated long-run cost functions that are poorly understood at this time. In the short run given capacities associated with low operating costs imply as primary short-run costs only line losses and congestion costs from capacity constraints. While these short-run costs are not directly incurred by the Transco, they need to be imposed upon the transmission users in order to induce them to use the transmission network efficiently. Any cost characterization presupposes a definition of outputs, which are themselves hard to characterize. We take primary outputs to be point-to-point transmission rights, with ancillary services as secondary outputs.

Demand for the primary transmission services is derived from the difference between electricity demands by LSEs and supply by generators for all node pairs. Provided an input demand is competitive it is usually easily derived from the vertical difference between demand downstream and the supply upstream. Due to the point-to-point nature of transmission, however, strong demand interactions exist between node pairs. Because of the homogeneity of electricity in the network the direct derivation of demand function can be complicated. It is hence best to estimate electricity demands at consumption nodes and generation supplies at each generation node and find the transmission demand as the solution to surplus maximization in the electricity markets between generators and LSEs.

This mimics perfect competition in market between generators and loads.[11] However, transmission demands may not be competitive, due to monopolies of LSEs and market power by generators. In this case the derived transmission demand may not be a well-defined function.[12] The problem of market power on the demand side would probably subside if demand for electricity were more price responsive (as a result of intelligent metering and pricing) or if electricity purchases by LSEs were covered by long-term contracts.

Vertical separation of generation, transmission and distribution is largely the result of regulatory reform of the electricity sector. Before the reforms of the 1990s it was presumed that vertical economies between these three production stages would dominate economies from competition under vertical separation. Competition in the electricity sector could be feasible under vertical integration of a (formerly) dominant electricity provider or under vertical divestiture of this provider. From the regulatory perspective these two cases differ in the approach to transmission price regulation. In both cases we may view transmission services as a bottleneck input that independent generators and LSEs require to trade electricity.[13] For such inputs the access pricing literature is relevant (Vogelsang, 2003 for telecommunications).

The access pricing literature differs distinctly between the case of vertical separation and that of vertical integration. Under vertical integration a firm would own and operate generation, transmission and distribution facilities and would compete with firms that only generate or market electricity. Such a (usually dominant) vertically integrated firm would, at the same time, supply transmission (and distribution) services as essential inputs to its unintegrated rivals. The countervailing interests as a competitor and input supplier have to be taken care of in any pricing rule and incentive scheme for such an integrated firm. The most relevant schemes in this case are access charges based on benchmark costs (TELRIC/TSLRIC in telecommunications, based on analytical cost models), the efficient component pricing rule (ECPR) and global price caps. Currently, vertical integration of this kind has been abolished or is being phased out in the U.S. electricity industry. Consequently, we are here concentrating on the case of vertical separation.[14] A vertically separated transmission providers could be a single dominant transmission company (Transco) or a set of separately owned merchant transmission providers. Operation and dispatch would in the latter case require a separate independent system operator (ISO), which may also be advantageous in case of a Transco.

[11] The ancillary products are essentially quality variables with public good characteristics. Thus, demands for ancillary products are hard to find empirically.

[12] In practice generators with market power may also bid a supply function that could be used for estimating transmission demand.

[13] Transmission is hard, but not impossible to substitute. Alternatives exist particularly in the long run.

[14] A more extensive application of the case of vertical integration is contained in Vogelsang (1999). While the advantages of vertical integration may be substantial it is not clear that they are preserved if a vertically integrated firm is forced to provide unlimited access to its bottleneck facilities. The case of vertical integration is highly relevant in the telecommunications sector. See the review of the issues in Vogelsang (2003).

Under vertical separation in the form of a Transco the multi-product monopoly regulation would be relevant. The Transco would ordinarily have no incentives to deny access to its network.[15] However, the Transco may want to charge profit-maximizing monopoly prices. In contrast, in the presence of economies of scale welfare-optimal prices would be lower in their level but their structure would not necessarily differ from those charged by an unregulated monopolist. There would have to be some markups on marginal costs that would vary inversely with the demand elasticities (Ramsey prices). If one interprets the fixed fee of a two-part tariff as the price for the right to use the transmission network the optimal Ramsey markup may be placed totally on this right-to-use, provided the demand for it is totally inelastic. In this case marginal cost pricing would become optimal, except for the fixed fee.

3.2 Suggestions for a PBR scheme for transmission pricing

A limited number of PBR schemes specifically for transmission pricing have been suggested in the literature. Among those, we here briefly discuss the merchant transmission approach, the Gans/King proposal (based on the ISS), the Léautier proposal and my own proposal.

Merchant transmission

The merchant transmission approach substitutes competition for regulation and is insofar outside the PBR proposals. However, merchant transmission has to be viewed as a serious substitute for the PBR approach and merchant transmission could benefit from the pricing formulas developed for PBR.

Merchant transmission requires separation of transmission ownership and control in order to allow for diverse ownership and common management (Joskow and Tirole, 2004). The literature shows that free entry by merchant transmission can lead to efficient transmission investment and overcome the monopoly problem of constraining capacity, provided a number of conditions are met. The main conditions include an absence of scale economies and of market power in the electricity wholesale market, the presence of well defined property rights to the stream of incremental congestion rents [positive and negative] generated by the investment and of sufficient complementary futures markets (Hogan, 1992; Bushnell and Stoft (1996, 1997). The planning of merchant transmission investments and the award of congestion rents further requires some regulation. Joskow and Tirole (2004) show that the efficiency of merchant investment breaks down under economies of scale/lumpiness[16] and under imperfect wholesale markets for electricity. They also show that gaming may occur between merchant investors and that merchant investment does not adapt well to uncertainty and the need for diversification.

[15] Such incentives could exist if incomplete bypass of the transmission network were feasible. Denying access to firms that also use bypass facilities could prevent such bypass.

[16] Joskow and Tirole show that in this case under efficient investment the revenues generated by congestion rents would be less than the costs of the capacity. One could now use a two-part tariff where the fixed fee would capture the remaining surplus form the additional capacity. However, this could lead to gaming in merchant investment because the additional surplus is likely to exceed the the trading loss from congestion pricing. Under free entry the expectation of positive profits would lead to gaming.

Furthermore, the interaction between expansion and maintenance (network deepening investments) is shown to be inefficient. While the Joskow/Tirole analysis is valid, their counterfactual is efficient investment behavior. That is a high standard that other proposals for monopolistic transmission investments by a Transco may also have a hard time to achieve. It will be particularly difficult to design Transco regulation that solves the problem of imperfect wholesale markets of electricity. Such regulation would be feasible under full information and unlimited transfers. However, if transfers are not allowed and if information is asymmetric (and if the relevant information is held by the – unregulated – generators) this market power problem can be solved only imperfectly. Thus, while merchant transmission may be highly imperfect, open entry into transmission investment may be a complementary tool to PBR.

Gans/King

Gans and King (1999) suggest using the ISS to induce efficient transmission investment. They do not formally introduce the ISS but simply note that the ISS would provide the firm with a reward (penalty) equal to the social surplus increase (decrease) in each period. The argument is then made through numerical examples. In particular, they claim that the ISS would also be capable of alleviating market power by generators efficiently. All this is based on the assumption that the relevant information is at the disposal of the regulator. Gans and King suggest that this information can be readily inferred from demand bids and generator bids. Furthermore, they suggest that these bids be used to construct counterfactuals without and with the investment. The financing of the ISS would then occur through "charging consumers the 'without-price' and paying generator the 'with-price'." The problem with this, as with the following approaches, is the question to what extent the short-run bidding behavior can be used as a guide for long-term investments.

Léautier/Nasser

Léautier (2000) proposes a regulatory contract that is closely related to the Gans/King proposal except that it is based in the Bayesian tradition. Léautier assumes that the incentive-pricing dichotomy holds so that pricing incentives can be separated from incentives for cost reductions. He then suggests offering the Transco a menu of revenue-sharing rules that trade off cost reductions against providing rents for the firm. There is no discussion of the practical implementation and financing of this part of the mechanism. Optimal transmission expansion is taken care of in the second part of the mechanism, according to which the Transco is "responsible for the full cost of the [operational] out-turn, plus the transmission losses, valued at the System Marginal Price." (Léautier, 2000, p. 77) The operational out-turn is "the sum over all generation nodes of the integral of the marginal generation cost minus the 'unconstrained' price, where the integral is taken between the constrained and 'unconstrained' generation." (p. 72) The operational out-turn is thus a measure of the cost of congestion to society. If the Transco is made to pay for all these losses it will minimize the sum of costs of transmission investment, transmission losses and social cost of congestion and will thereby be induced to invest efficiently. Calculation of the operational out-turn is similar to that of the surplus change under the ISS proposed by Gans and King. In Nasser (1997, pp. 206-209) Léautier also proposes a non-Bayesian scheme related to Finsinger and Vogelsang

(1982). In this case, instead of any other revenues (that go to the regulator) the Transco, in each period t, receives a subsidy equal to the investment payment plus a fraction of the "net corrected revenue." Here the net corrected revenue is the sum of two components. The first is the sum of the value of flowgate rights multiplied by the net change in line capacity between t-1 and in t. This component can be interpreted as the Slutsky approximation of the consumer surplus change from the capacity expansion. The second component is a fraction of the variable costs of investment in period t. In a stationary environment, this scheme would (like Vogelsang (2001) below) converge to optimal capacities over time. The problem of the Léautier/Nasser proposals is, however, that they are silent about how the costs borne by the Transco will be financed by the regulator.

Vogelsang and Rosellón/Vogelsang

While the Gans/King and Léautier/Nasser proposals explicitly calculate the optimal regulatory scheme for each transmission investment, the proposal by Vogelsang (2001) is based on the price-cap tradition. Under Vogelsang (2001) the sequence of moves would be

(1) The regulator sets the regulatory pricing constraint (given by equation 7 below).
(2) The Transco collects information about generation supply and electricity demand at all relevant geographical locations (or at each node).
(3) The Transco invests in grid capacity.
(4) The Transco sets point-to-point transmission prices.
(5) Generators and loads sign bilateral electricity contracts and buy point-to-point transmission services.
(6) There can be excess supply or excess demand for transmission services on a point-to-point basis. Excess supply could hurt the Transco but would not cause any feasibility problems. Excess demand could cause feasibility problems (although, together with excess supply for other point-to-point relationships the total sum might still be feasible). The Transco could then use non-price rationing and sell point-to-point transmission services to bilaterals on a first-come-first-serve basis. Regulators could impose penalties, giving the Transco incentives to price in such a way that excess demand does not occur (creating excess supply).
(7) The Transco calculates the fixed fee from the regulatory constraint and charges it to the loads.

Alternatively, moves (4)-(7) could be replaced by:

(4a) There is an ISO, who asks for (sequences of) bids from generators and loads at each node and then calculates nodal prices. Loads (ex post) pay the ISO according to their last bids and generators receive payment of their last bids in such a way that markets always clear. The Transco receives as congestion payments the difference between what loads pay and what generators receive. Fixed fees are then calculated from the regulatory constraint and are paid by the loads. In this case the Transco does not set prices but only makes available capacities.

Vogelsang (2001, in the formulation of Rosellón and Vogelsang, 2004) assume that, for a load-flow model for real power (DC load approximation), the Transco's profit maximization problem is given by:

$$\underset{p^t, F^t}{Max}\, \pi^t = p^t q^t + F^t N^w - c(K^t, H^t) \tag{4}$$

s. t.

$$\iota^t q^t = 0 \quad \forall t \tag{5}$$

$$H^t(Z^t, graph^t)\tilde{q}^t \leq K^t \quad \forall t \tag{6}$$

$$p^t q^w + F^t N^w \leq p^{t-1} q^w + F^{t-1} N^w \quad \forall t \tag{7}$$

where:

p^t = price vector of transmission services in period t

q^t = the vector of transmission services in period t ($\sum_j \gamma_j^t = [q^t]; \gamma_j^t = \begin{bmatrix} - & x \\ 0 & \\ 0 & \\ \cdot & \\ \cdot & \\ + & x \\ 0 & \end{bmatrix}$)

$\tilde{q}^t = [q_{ij}^t]e^t$ in (6) are net injections rather than point-to-point transmission services. Thus, we get the vector of net injections by multiplying the matrix of point-to-point transmission services by a unit vector.

F^t = fixed fee in period t

N^w = weighted number of consumers

$c(K^t, Z^t)$ = present cost of capacity

ι^t, e^t = vectors of ones

H^t = transfer admittance matrix at period t

Z^t = vector of impedances at period t

$graph^t$ = topology of the network at period t

K^t = available transmission capacity in period t

If the weights of the price-cap mechanism are chained Laspeyres weights, if the transmission cost function is well-behaved, if the environment does not change and if

transmission capacity can be adapted in every period the mechanism converges to transmission capacity and prices that are surplus maximizing.[17] With idealized weights the mechanism converges in a single period. Because of a changing environment and because transmission investments are likely to be lumpy, an approximation of idealized weights would therefore be crucial for the mechanism to become practically relevant.

It is challenging to regulate fluctuating prices while keeping the long-run incentives. Price caps can be used for that purpose because they provide stable price levels along with flexibility of the price structure. As shown in Vogelsang (2001) such combination can be achieved in price caps with two-part tariffs, where the fixed part compensates for variations in the variable parts. Then both good capacity expansion and capacity utilization properties are achievable. Vogelsang (2001) is the only mechanism so far that directly links real-time pricing with investments. However, Vogelsang shows that in the case of real-time pricing an ISO would be required to ensure that those prices truly reflect the network scarcity at any point in time. Otherwise, the Transco would have incentives to distort real-time pricing under the mechanism.

Conclusions from the literature

The suggestions made in the literature so far are not well adapted to the length of life of transmission investments and the environmental changes associated with that length. The Gans/King and Léautier proposals both base the investment incentives on the bid curves of generators and loads but do not specify the periods of observation and how the regulator would aggregate this information into a forecast that would be relevant for long-term investment. The Vogelsang proposal shifts this problem from the regulator to the Transco that is made responsible for forecasting correctly. However, neither of the approaches deals with the credibility of the regulator in the long run.

3.3 A synthesis approach based on a three-period framework

In my view, the main problem with the PBR schemes for transmission proposed in the literature is their failure to come to grips with the specific timing issues of transmission services and networks. We therefore explicitly introduce a framework based on those timing issues. This framework is based on the distinction between ultra-short, short and long periods.

3.3.1 The framework

The ultra-short period

The ultra-short period is motivated by the desirability of real time pricing, which is a distinct and differentiating feature of transmission pricing. Real-time pricing could also be substituted by peak-load pricing.[18] During the ultra-short period the Transco (or ISO) makes dispatch decisions only (incl. demand curtailment) and there are no meaningful

[17] In principle, this pricing mechanism can be extended to merchant investors, who each would receive fixed fees with precisely the same property.

[18] The advantages and disadvantages of real-time pricing vs. peak-load pricing are reminiscent of those between point-to-point vs. zonal pricing.

possibilities for transmission cost reductions. Thus, this is only an allocative efficiency problem. Pricing occurs largely under certainty about supply and demand. Demands for transmission services will fluctuate wildly between ultra-short periods. The demand elasticity for transmission services will be determined mostly by electricity supply because electricity demand is very inelastic in the very short run. Regarding PBR, full regulatory commitment can be expected. Steep incentives would be adequate because the relevant activity does not extend beyond the period of commitment and the information is also not valid beyond.

The short period

The short period coincides in terms of pricing with the application of RPI-X for each of these periods. It is also the period for calculation of fixed fees.[19] During such a short period decisions are made on operations, repairs and maintenance. The ability of the Transco to influence costs can be limited during this period. Thus, allocative efficiency would be most important, but productive efficiency can also play a role. Pricing occurs under some market uncertainty. Demands for transmission services will be more elastic than in the ultra-short periods because there will be some adjustments by electricity customers in the use of appliances. For each of these short periods regulatory commitment is complete, and there is almost full regulatory commitment over several of them. For activities restricted to this period and with parallel myopic information steep incentives are in order.

The long period

In terms of pricing the long period is given by the time between revisions of the PBR mechanism applied. It would, for example, apply to a change in RPI-X formula. The long period could coincide with the length of the regulatory contract. This period is too short to cover network investment decisions, which extend over several such periods (time to build plus life of asset plus life of overlapping assets). Also, the time for the validity of information about investment is likely to be similarly long. Thus, the motivation for the length of this period is not given by the time horizon of investment but rather by the ability to commit, given external changes over time. PBR regulation is often seen as applicable only in the short run. If regulatory commitment to a PBR cannot be guaranteed beyond a single long period and if investments last over several or many of such periods, how can PBR aid investment? Does the failure of commitment mean that PBR also would not work during each long period if there were more than one? We interpret the absence of regulatory commitment to a PBR mechanism not to be the absence of any commitment. Rather, it means the absence of commitment beyond some basic rules (such as the restrictions implied by the 1944 Hope decision of the US Supreme Court). These basic rules are implementable because they are based on the U.S. Constitution and on fairness principles. As indicated above, rate-of-return regulation with a "used and useful" criterion could be the basis for PBR revisions after long periods. The interaction of competition (merchant investment) and dominant firm regulation could also play a role for the long-term revisions. As the Transco can definitely influence its costs during the

[19] If fixed fees and RPI-X apply to different period lengths (e.g., month vs. year) the short period could actually be disaggregated further. This could play a role if the mechanism used does not converge in a single period.

period and over several of such periods, incentives for cost reductions would be desirable. However, all one can hope for is (a) that the period between reviews can be extended sufficiently (with commitment) under flexible PBR that traces the firm's costs well enough to mimic the outcome of rate-of-return regulation during that time, (b) commitment can be sustained longer than a single long period under repeated interaction, with regulatory reputation at stake and new investments required in a growing market. Between long periods the demands for transmission services will also be affected by location decisions of electricity generators. The interaction between such generation investments and transmission investments would be part of the evaluation of "used and useful" investments under transmission regulation.

Linking the periods

Performance-based regulation has increased pricing flexibility in some areas and thereby been conducive to liberalization if only as a concession to make incumbents willing to accept sector reforms. The impact of amending or replacing rate-of-return regulation by incentive regulation (price caps) increases short-term efficiency incentives. However, rate-of-return regulation had been a commitment device against expropriation and therefore has shielded sunk, lumpy investments (Newbery, 2000). The 1970s and 1980s showed that this device no longer worked. Rate-of-return regulation prevented innovative investment and was associated with the nuclear cost overruns. Its further problem is that it is not compatible with competition because it reduces the incumbent's cost-reducing incentives and its pricing flexibility. Thus, regulation has to combine flexibility of PBR with assurances of rate-of-return regulation.

While the range between the ultra-short periods and a single long period is linked by credible commitment, the long periods themselves are not. The Bayesian literature would therefore suggest that the incentives within a short period might have to be softened, compared to a regulatory problem that would consist of only a single long period. However, this depends on (a) the types of decisions to be made and the knowledge gained by the regulator during the types of periods and (b) the consequences of lack of commitment.

Clearly, network expansion investments carry over several long periods so that for them the lack of commitment would be relevant. The question, however, is if information from spot pricing and from short-term O&M decisions carries over beyond a single long period. I conjecture that it does. For example, the spot-price information can and will be used to inform the network expansion and maintenance decisions. Similarly, the maintenance information will be used for investment decisions and for future maintenance decisions. Also, maintenance decisions influence the replacement investments. Thus, while the informational link between ultra-short, short and long periods may not be as strong as the links in the shorter run, it is still existent and potentially relevant for incentives. This suggests that profit sharing may be preferable to a strict RPI-X regime for the link between the short periods. Nevertheless, previous experience with profit sharing that broke down under inflation suggests that the RPI-X formula should be combined with profit sharing.

The important question is therefore, what is the meaning of the non-commitment beyond a single long period. This non-commitment means something very different in the U.S. than in other countries without the rate-of-return regulation tradition.[20] In the U.S. it means that regulated firms can only expect a competitive rate of return on their investments and that they can expect such a return only if their investments are found to be prudent ("used and useful"). At the same time, they cannot expect to receive a higher rate of return. In my view, this introduces a potential bias that is contrary to the assumption of the Bayesian literature (and more in line with Joskow, 1974). Under the Bayesian incentive regulation approach with lack of commitment, the firm is expected to be penalized by a downward price adjustment if it does well (has low costs). In contrast, under U.S. style rate-of-return regulation the regulated firm would be penalized if it does badly (overestimates demands and/or underestimates costs). Whether this potential penalty is responsible for the observation that allowed rates of return are higher than the cost of capital is an open question. What does this imply for short-term incentive schemes in combination with rate-of-return regulation? Provided rate-of-return regulation is upheld in the long run, it implies that the firm can expect to have part of any gain (or loss) from short-term incentives in excess of the allowed rate of return be taken away ex nunc, after a long-term regulatory review.[21] The effect of this will depend on the extent of achievable gains relative to the overall rate of return of the firm.

Thus, while rate-of-return regulation provides for some commitment, even over several or even many long periods, the question is if there can be other long-run regulatory variables that are constant (or consistent) over the long run. For example, can the firm expect that the types of regulatory constraint, such as real-time pricing with two-part tariffs (with adjustments to the fixed fees) or the RPI-X formula (with adjustments of the X-factor), will remain in the long run? That is obviously not the case. As a result regulated firms will not make optimal investments in their full adaptation to the features of the mechanism.

3.3.2 The approach to each type of period

I am going to base the arguments in the following on Vogelsang (2001). Recall that in that mechanism investment (and disinvestment) occurred at the beginning of each period, followed by variable pricing that would coincide with actual transmission, while fixed fees would be calculated at the end of the period. Thus, the mechanism lumps together the short period and the long period and assumes that investments do not extend beyond a single such period. In Rosellón and Vogelsang (2004) we try to make investments last beyond single periods and do not allow for disinvestments. However, the three-period approach outlined in the last section goes substantially beyond that. Under the three-period approach, no investment might occur for many short periods and even for times beyond a long period. The influence of the mechanism on investment would only occur via the long-run rate-of-return constraint in conjunction with price predictions from the operations of the mechanism in the past.

[20] For this issue of country-specific institutional endowments, see Levy and Spiller (1996).
[21] To the best of my knowledge no important court decisions have been made to the contrary.

The ultra-short period

Since the ultra-short period is only concerned with pricing and since there is full commitment, a fully efficient mechanism can be used if available. In Vogelsang (2001) I argue that real-time pricing is not compatible with an index approach to price regulation. Thus, if one wants to use price caps based on a price index, peak-load pricing with identifiable ultra-short periods would be required. In contrast, real-time pricing would work only under average revenue (AR) constraints, where for each point-to-point transmission service the total revenues from real-time pricing would be summed over all ultra-short periods in a short period and then divided by the number of ultra-short periods. Because such AR constraints would be soft on a Transco's pricing, this approach may require the aid of an ISO, who makes sure that the real-time prices are the actual congestion prices. Real-time congestion pricing would be optimal but not cost covering if there are economies of scale or stranded costs that need coverage. This would either suggest cost coverage through a not fully efficient peak-load pricing mechanism or eliminating the deficit from congestion pricing by charging a fixed fee. Such fixed fees would only be calculated and billed over longer periods, linking the ultra-short period with the short period.

The short period

Pricing in the short period is characterized by the collection of the fixed fee and by some adjustment according to formula. This adjustment could be of the RPI-X type, the benchmark (= yardstick) or the profit-sharing type with an inflation adjustment. In either case, the expectation is that the regulated firm, if operating efficiently, can at least cover its costs and can possibly do better. This period therefore requires regulatory commitment for several such periods. As a result, prices can be allocatively efficient, while incentives for short- to medium-term managerial and operating efficiency are provided.[22]

The long period

Regulatory commitment is assumed not to last beyond a long period, which covers several short periods. At the same time, network (expansion) investments last over several long periods. The Bayesian literature therefore suggests that investment-inducing incentives should be kept low-powered, such as under rate-of-return regulation. At the same time, the literature on rate-of-return regulation has gone beyond the Averch-Johnson effect by including the "used and useful" criterion as a means of inducing efficient investment levels. Thus, we revert to this device for the long-run revisions of the PBR.

This would be relevant for the starting prices under PBR. Such prices should definitely cover forward-looking costs. However, they may also have to contain a surcharge on forward-looking costs to finance stranded assets (defined as the amount by which bookkeeping values of assets exceed their economic value). Often, at the beginning of PBR the starting prices can be taken to be those under cost-of-service regulation and any adjustment to the desired level would be achieved through the X-factor or profit sharing.

[22] The incentives obviously lessen toward the end of each long period. In principle, the regulator can deal with this problem through appropriate time-specific weighting. See Vogelsang (1989).

When it comes to revisions at the end of a long period, a one-time adjustment can be made to reach prices that cover forward-looking costs (assuming that stranded costs have vanished over time) or the X-factor (or increased profit sharing) can be used to get there gradually. Further adjustments can be made for the consequences of lumpy investments. Because of the "used and useful" criterion those investment costs would not prudently be added on a forward-looking basis (which would otherwise be economically correct) but only when used. Since under PBR for a Transco all these investment costs would be covered by adjustments of the fixed fees, the allocative distortions from such a postponement of cost reimbursement (including interest) would be minimal.

A combined approach for all types of periods

A combined approach could be based on a blend between Vogelsang (2001) and the ISS with a verifiable approximation of consumer surplus change and financing of the 'subsidy' through the fixed fee of a two-part tariff. Recall that under the approximated ISS the fixed fee in any period t would be

$$F^tN = ½(p^{t-1} - p^t)(q^t + q^{t-1}) - \pi^{t-1} \qquad (8)$$

Here π_{t-1} is the profit (excess rate of return over cost of capital) generated last period from the nodal prices only. Sappington and Sibley show that pricing converges in a single period to the efficient level and stays there, provided cost and demand parameters do not change.[23] It is easy to show that this extends to Transco investment. In particular, the Transco would receive a total economic profit for any period in which it improves efficiency by investment (in expansion). The profit would equal the efficiency increase. In each period after that the firm would be reimbursed any losses incurred under nodal prices. This means that the fixed fees thereafter (in periods without transmission expansion) simply assure that the firm covers all its costs after deficits from nodal prices.

In contrast to (8) under Vogelsang (2001) the regulatory constraint with the approximation to consumer surplus change would be

$$F^tN = F^{t-1}N + ½(p^{t-1} - p^t)(q^t + q^{t-1}) \qquad (9)$$

There are two main differences between (8) and (9). First, (9) provides for cumulative consumer surplus increases to the Transco. Second, in (8) last period's profit is deducted. Now, we can combine the two mechanisms by using (9) for pricing in the ultra-short and short periods,[24] an RPI-X adjustment or profit sharing for the short periods and a profit adjustment (by total excess profits rather than excess profits from nodal prices only) at the end of each long period. This means that prices p^t would be average revenues from the ultra-short periods τ: $p^t = \Sigma p^\tau q^\tau / \Sigma q^\tau$. The RPI-X and the profit sharing adjustments would only change the fixed fees and have no effects on efficient variable prices. These adjustments would, however, partially counteract any consumer surplus increases that are handed to the Transco cumulatively. If this interplay is anticipated well by the regulator's

[23] However, it is not clear how the properties extend to (a) the general multiproduct cases with interdependent demands and (b) shifts in demands.

[24] For the combination of short and ultra-short periods see Vogelsang (2001), pp. 155-160.

choice of X and/or the profit-sharing parameter s the long period of commitment could be extended somewhat. Biglaiser and Riordan (2000) show that if the long period between price cap reviews is too short then this can lead to underinvestment. Thus, extending this period is likely to be beneficial (although, according to Biglaiser and Riordan, the period can also be too long).

4. Conclusions and open research questions

4.1 Conclusions

Economists like to optimize. Performance-based regulation (PBR) should be optimal. What does that mean? Until about 1970, many of us believed in truly optimal or first best regulation, which meant marginal cost prices. However, over time, the adjective "optimal" has received more and more qualifications (Vogelsang, 1999). The first was that losses incurred under optimal prices in the presence of economies of scale led to second best Ramsey pricing, a movement that peaked around 1980. The main insight here was that prices should deviate from marginal cost prices by markups that are inversely proportional to demand elasticities (or, more precisely, to super-elasticities). The deficiency of Ramsey prices was the regulator's lack of information about cost and demand functions. Thus, the next wave was third best regulation under incomplete information. The main insights from this wave were (a) that regulated firms might need to be able to make economic profits in order to reveal private information and (b) that such profits can be limited by giving firms a choice from a menu of regulatory options. This wave probably peaked with the publication of the Laffont and Tirole (1993) book on incentive regulation. What is the next step away from optimal price regulation? Is it fourth best regulation that makes theoretical models of regulation applicable under political and practicality constraints? In any case, regulation economists have moved further and further away from what was once perceived as optimal price regulation. Consequently, in order to be relevant, the price regulation mechanisms we consider here are not strictly optimal in that they maximize a well-defined social welfare function. Rather, the PBR schemes are meant for practical application and thus should have some desirable properties.

A specific list of such objectives for transmission pricing has been developed by Green (1997). According to him transmission pricing should fulfill six sensible principles. They are

1. Efficient day-to-day operation of the bulk power market
2. Efficient investment in the transmission system
3. Signaling of locational advantages for generation and distribution investments
4. (Historic) cost recovery of transmission assets
5. Simplicity and transparency
6. Political feasibility.

Among Green's (1997) principles for transmission performance we have concentrated our discussion on the functioning of the bulk power market, on cost recovery and on transmission investment.

Principles 1 (a functioning bulk power market) and 2 (transmission investments) would be assured under our mechanism, provided the generators have no market power. To the extent that the transmission users have market power transmission pricing may need to counteract this downstream market power by pricing below marginal costs. This would compensate for but does not necessarily eliminate that market power. It would be associated by larger transmission capacity.

In order to achieve principle 3 (optimal location of generation capacity), the Transco would either have to set predictable variable fees and fixed fees that directly relate to transmission capacity costs caused by new generation/distribution capacity or would have to engage in long-term contracting with its customers. Both of these are feasible under the proposed scheme but not automatic parts of it. The suggested regulatory approach is definitely compatible with stable average prices incurred by customers that would signal transmission investment costs. In addition, the approach can also be implemented through contracts (as options or as part of tariffs). For example, customers could buy interruptible service at lower fixed fees or firm services at higher fixed fees (and, possibly lower or vanishing variable fees). All this could be done within the price-cap constraint. However, since transmission users would have to pay fixed fees that are independent of their individual usage and since most of these users would remain transmission users independent of their generation investments, potential generation investors would only be able to use the variable transmission prices as signals for their generation investments. This could lead to insufficient generation investments that would be substitutes for transmission capacity.

Principle 4 (cost recovery) can be achieved through initial rates that reflect embedded investment costs. In this case, the 'X' factor or profit sharing would have to account for the difference between embedded average costs and forward-looking incremental costs. If embedded average costs exceed forward looking incremental costs 'X' or the profit share going to consumers should be positive, inducing lower prices and forcing the firm to reduce its costs through investment. Vice versa, if embedded average costs are below forward-looking incremental costs, 'X' or the "profit" share going to consumers should be negative and induce higher prices. In the long run, adjustments according to rate-of-return regulation with a "used and useful" criterion would assure cost recovery.

Principle 5 (simplicity and transparency) is in the eyes of the beholder. Clearly, the regulatory mechanism has to be based on transparent data. The level of complexity of actual tariffs depends on the tradeoff between efficiency and complexity that market participants and regulators are willing to make. Since participants in the transmission market are largely sophisticated firms, simplicity would have less value here than in the retail market for electricity.

Principle 6 (political feasibility) requires that no interest group involved is made noticeably worse off. It is closely linked with principles 4 and 5. Principle 4 assures that the Transco is not made worse off. In addition, basing initial rates on historic costs and choosing 'X' carefully assures that generators, industrial users and distribution companies receive services on average at better than status-quo prices. However, that does not necessarily mean that all of them are better off. First, better transmission can intensify competition between generators, thus reducing profits of some of them. Second, more sophisticated pricing means that former cross-subsidies may be eliminated.

4.2 Open research questions

Several aspects of the types of PBR discussed above require further research.

While the regulator and Transco may gather valuable information about the profitability of transmission investments from the short-term bids of generators and load-serving entities or from the fluctuation in congestion prices, it is not obvious how this short-term information translates into projections necessary for long-term investments. Obviously, long-term demands for transmission services depend on lumpy investments by generators and loads and potentially on changes in consumption behavior induced by improved consumption metering and smart scheduling of appliances. It would be highly valuable to generate such long-term information.

A related research question is how long the half-life of information is that the regulator gains from the revealed behavior of the Transco. The shorter this half-life is, the less would a lack of regulatory commitment matter. This half-life may vary for different activities, such as the costs of expansion investments (where it may be very long) as opposed to the costs of operations and maintenance (where it may be much shorter).

Ever since the California electricity disaster the market power of generators has become a major regulatory nightmare in states that have reformed their electricity sectors. Increased transmission capacity has been suggested as a means of reducing such market power. It would therefore be important to research the effects of PBR for Transco investments and their interaction with (a) generation investments and (b) short-term generator behavior.

We have used a linear approximation to the consumer surplus change in equations (8) and (9) above. If the approximation is (almost) correct it would provide full and immediate incentives for efficient investment levels. However, it is not clear that the approximation is correct if (a) demands for the different point-to-point transmission services are interdependent and if (b) demands shift over time. Demand interdependence in transmission systems is very likely. However, it may be symmetric so that the integrability conditions for the demand system hold. We will try to determine if those conditions are sufficient for the approximation to be correct. Exogenous demand shifts would definitely bias the approximation because part of the quantity changes would be due to demand shifts rather than to price changes. Thus, the approximation would have to be adjusted for such demand shifts. We will try to find ways to make such adjustments.

Fixed fees are usually a good way to raise revenues without generating severe allocative distortions. However, if most of the revenues are generated from fixed fees, users may want to avoid such fees and fairness questions arise if small or intermittent users have to pay the same fixed fees as large continuous users. Vogelsang (2001) discusses some methods for calculating fixed fees in a more discriminating fashion. However, the best basis for fixed fees has yet to be determined. This could also take care of the issue of demand shifts.

We have not considered quality aspects, reliability in particular. Our incentive regulation proposal relies largely on the Transco's profit incentive to provide quality of service. Bad quality in the form of congestion, for example, would either lead to high variable fees that would be penalized through lower fixed fees. Or it would lead to foregone sales. However, the price-cap scheme may not take care of all quality dimensions and, to the extent that price caps constrain profits, there may exist incentives to reduce costs by reducing some quality attributes. To prevent this, quality incentives, standards and commitments may need to be added to the regulatory scheme, for example, in order to prevent poor ancillary services and outages. Quality parameters that may require attention include system reliability, customer service or employee safety. The problem of valuing quality poses a big problem for the implementation of PBR.[25]

We have discussed merchant entry only as an alternative to PBR of Transcos. However, allowing merchant entry may also be a complementary instrument. Allowing entry is usually a credible long-term commitment because entrants quickly form a strong interest group countervailing the political clout of the incumbent. Merchant entry may therefore act as a lid to incumbent inefficiency in investment. However, the question is if there could be too much merchant investment (due to loop flows).

References

Armstrong, Mark and David Sappington (2003), "Recent Developments in the Theory of Regulation," forthcoming in M. Armstrong and D. Sappington (eds.), *Handbook of Industrial Organization*, Vol. III, Amsterdam: North Holland.

Baron, David and David Besanko (1984), "Regulation, Asymmetric Information and Auditing," *RAND Journal of Economics* 15, pp. 447-470.

Baron, David and Roger Myerson (1982), "Regulating a Monopolist with Unknown Costs", *Econometrica* 50, pp. 911-930.

Baumol, William J. (1967), "Reasonable Rules for Rate Regulation: Plausible Polices for an Imperfect World," in A. Phillips and O. E. Williamson (eds.), *Prices: Issues in Theory, Practice, and Public Policy*, Philadelphia: University of Pennsylvania Press, pp. 108-123.

Baumol, William J. (1982), "Productivity Adjustment Clauses and Rate Adjustment for Inflation," *Public Utilities Fortnightly*, July 22, pp. 11-18.

[25] See Navarro (1996), p. 141/142, who proposes, among others, hedonic pricing or contingent valuation.

Beesley, Michael and Stephen Littlechild (1989), "The Regulation of Privatized Monopolies in the United Kingdom," *Rand Journal of Economics* 20, pp. 454-472.

Biglaiser, Gary and Michael Riordan (2000), "Dynamics of Price Regulation," *RAND Journal of Economics* 31, pp.744-767.

Boiteux, Marcel (1956), "Sur la Gestion des Monopoles Publics Astreints à l'Equilibre Budgetaire," *Econometrica* 24, pp. 22-40; translated as "On the Management of Public Monopolies Subject to Budgetary Constraints," *Journal of Economic Theory* 3, 1971, pp. 219-240.

Brown, Lorenzo, Michael Einhorn and Ingo Vogelsang (1989), "Incentive Regulation: A Research Report," Federal Energy Regulatory Commission, Office of Economic Policy, Technical Report 89-3, November.

Brown, Lorenzo, Michael Einhorn and Ingo Vogelsang (1991), "Toward Improved and Practical Incentive Regulation," *Journal of Regulatory Economics*, 3, pp. 323-338.

Bushnell, J. B. and S. E. Stoft (1996), "Electric Grid Investment Under a Contract Network Regime," *Journal of Regulatory Economics* 10, pp. 61-79.

Bushnell, J. B., and S. E. Stoft (1997) "Improving Private Incentives for Electric Grid Investment," *Resource and Energy Economics* 19, pp. 85-108.

Crew, Michael and Paul Kleindorfer (1996), "Incentive Regulation in the United Kingdom and the United States: Some Lessons," *Journal of Regulatory Economics* 9 pp. 211-225.

Evans, Lew and Steve Garber (1988), "Public Utility Regulators Are Only Human: A Positive Theory of Regulatory Constraints," *American Economic Review* 78, pp. 444-462.

Finsinger, Jörg and Ingo Vogelsang (1982), "Performance Indices for Public Enterprises," in L.P. Jones (ed.): *Public Enterprise in Less Developed Countries*, New York: Cambridge University Press, 1982, pp. 281-296.

Gans, Joshua S. and Stephen P. King (1999), "Options for Electricity Transmission Regulation in Australia," University of Melbourne, September 10.

Gilbert, Richard and David M. Newbery (1994), "The Dynamic Efficiency of Regulatory Constitutions," *RAND Journal of Economics* 25, pp. 538-554.

Green, Richard (1997), "Electricity Transmission Pricing: An International Comparison," *Utilities Policy* 6, pp.177-184.

Hogan, William F. (1992), "Contract Networks for Electric Power Transmission," *Journal of Regulatory Economics* 4, pp. 211-242.

Hotelling, Harold (1938), "The General Welfare in Relation to Problems of Taxation and Railway and Utility Rates," *Econometrica* 16, pp. 242-269.

Joskow, Paul L., (1974), "Inflation and Environmental Concern: Structural Change in the Process of Public Utility Price Regulation," *Journal of Law and Economics* 17, pp. 291-327.

Joskow, Paul L. and Paul W. MacAvoy (1975), "Regulation and the Financial Conditions of the Electric Power Companies in the 1970s," *American Economic Review* 65, pp. 295-311.

Joskow, Paul L. and Richard Schmalensee (1986), "Incentive Regulation for Electric Utilities," *Yale Journal on Regulation* 4, pp. 1-49.

Joskow, Paul and Jean Tirole (2004), "Merchant Transmission Investment," March 31.
Laffont, Jean-Jacques and Jean Tirole (1986), "Using Cost Observation to Regulate Firms," *Journal of Political Economy* 94, pp. 614-641.
Laffont, Jean-Jacques and Jean Tirole (1993), *A Theory of Incentives in Procurement and Regulation*. Cambridge, MA: The MIT Press.
Léautier, Thomas-Olivier (2000) "Regulation of an Electric Power Transmission Company," *The Energy Journal* 21, pp. 61-92.
Levy, Brian and Pablo T. Spiller (eds.) (1996), *Regulation, Institutions, and Commitment*, Cambridge, UK, New York and Melbourne: Cambridge University Press.
Lewis, Tracy and David Sappington (1990), "Sequential Regulatory Oversight," *Journal of Regulatory Economics* 2, pp. 327-348.
Littlechild, Stephen C. (1983), *Regulation of British Telecommunications' Profitability*, Report to the Secretary of State, Department of Industry, London: Her Majesty's Stationery Office.
Loeb, Martin, and Wesley A. Magat (1979), "A Decentralized Method of Utility Regulation," *Journal of Law and Economics* 22, pp.399-404.
Nasser, Thomas-Olivier (1997), "Imperfect Markets for Power: Competition and Residual Regulation in the Electricity Industry," Unpublished Ph.D. Dissertation, MIT.
Navarro, Peter (1996), "The Simple Analytics of Performance-Based Ratemaking: A Guide for the PBR Regulator," *Yale Journal on Regulation* 13, pp. 105-161.
Newbery, David M. (2000), *Privatization, Restructuring, and Regulation of Network Industries*, Cambridge, MA, and London: MIT Press.
Sappington, David and David Sibley (1988), "Regulating Without Cost Information: The Incremental Surplus Subsidy Scheme," *International Economic Review* 29, pp. 297-306.
Vogelsang, Ingo (1988), "A Little Paradox in the Design of Regulatory Mechanisms," *International Economic Review* 29, pp. 467-476.
Vogelsang, Ingo (1989), "Price Cap Regulation of Telecommunications Services: A Long-Run Approach", in M.A. Crew (ed.), *Deregulation and Diversification of Utilities*, Boston: Kluwer Academic Publishers, pp. 21-42.
Vogelsang, Ingo (1999), "Optimal Price Regulation for Natural and Legal Monopolies," *Economia Mexicana* 8, pp.5-43.
Vogelsang, Ingo (2001), "Price Regulation for Independent Transmission Companies," *Journal of Regulatory Economics* 20 pp. 141-165.
Vogelsang, Ingo (2002), "Incentive Regulation and Competition in Public Utility Markets: A 20-Year Perspective," *Journal of Regulatory Economics* 22, pp. 5-28.
Vogelsang, Ingo (2003), "Price Regulation of Access to Telecommunications Networks," *Journal of Economic Literature*, XLI, September, pp. 830-862.
Vogelsang, Ingo, and J. Finsinger (1979), "A Regulatory Adjustment Process for Optimal Pricing by Multiproduct Monopoly Firms," *Bell Journal of Economics* 10, pp. 157-171.