Optimal Price Regulation for Natural and Legal Monopolies

Ingo Vogelsang*

Abstract: Optimal price regulation for natural and legal monopolies is an impossible task. The still difficult task of good price regulation can be systematized by considering separately price level and price structure of the regulated firm. Various methods of price level and price structure regulation are evaluated and then considered for the regulation of electricity transmission, both in the context of an independent transmission company and of vertical integration between transmission and most of the generation capacity. The regulatory approach suggested uses price caps defined on two-part tariffs. This way, flexibility for short-term capacity utilization can be combined with incentives for investments in new transmission capacity.

Resumen: La regulación óptima de precios para monopolios naturales y legales es una tarea imposible. La aún difícil tarea de una adecuada regulación del precio puede ser sistematizada considerando separadamente el nivel y la estructura del precio de la empresa regulada. Se evalúan diversos métodos de regulación del nivel y la estructura de precios y posteriormente se les considera para regular la transmisión de electricidad tanto en el contexto de una compañía de transmisión independiente como en el de la integración vertical entre transmisión y la mayor parte de la capacidad de generación. El enfoque regulatorio sugerido utiliza precios máximos definidos como tarifas en dos partes. De esta forma, la flexibilidad para la utilización de la capacidad de corto plazo puede combinarse con incentivos para inversiones en nueva capacidad de transmisión.

* Boston University. Paper prepared for CIDE Seminar on Structural Reform and Regulation in the Energy Sector, Mexico City, August 20/21, 1998. The author would like to thank Juan Rosellón for helpful comments.

1. Introduction

In a way, the title of this paper is an anachronism. Consider the first three words, "optimal price regulation". 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. The first was that losses incurred under optimal prices in the presence of economies of scale led to *second best* Ramsey pricing, which 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 that regulated firms might need to be able to make economic profits in order to reveal private information and 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 schemes are meant for practical application and thus should have some desirable properties. The desirable properties are two in particular. First, in order to be implementable, price regulation should be viewed as fair, meaning that it makes no important group significantly worse off than the *status quo*. Second, and closely related, there has to be an efficiency improvement.[1]

Now consider the remainder of the title, "natural and legal monopolies". For the last few years, the decline of *legal monopolies* has been a worldwide phenomenon. Competition (often accompanied by privatization) has been introduced in traditional monopoly sectors, such as telecommunications and electricity supply. There has been a similar decline in the reported incidence of *natural monopoly* in these sectors. Only niches of natural and, in some cases, legal monopolies remain. Examples could include local telephone companies, gas distribution companies, electricity distribution and transmission companies. In spite of this decline, why might it nevertheless be worthwhile to cast a new look at the regulation of these monopolies in light of limits to optimal regulation? All the above examples have in common that they belong to industries with potentially competitive parts that are vertically related to these monopolies. In the energy sector, electricity transmission and distribution have strong monopoly attributes, while generation is potentially competitive. Natural gas and oil pipelines and gas distribution companies are potential monopolies, while gas and oil production could be competitive. Thus, monopoly regulation today almost inevitably has to be seen in the context of related competitive industries. Furthermore, the monopolists often are vertically integrated so that they either hold a monopoly over all stages or compete with firms that use the monopoly outputs as their inputs.

In the following, we will first treat the monopoly price regulation problem in a generic fashion, separating the problems of price level and price structure. Then we will concentrate on energy industries in a mixed monopoly/competitive setting, first without then with vertical integration. The paper takes electricity transmission as the main example, because it raises the most complex issues. It is fairly straightforward to move from here to natural gas and oil pipelines, or to electricity and gas distribution companies.

2. Price Level Regulation

Price regulation is best discussed separately for the level and structure of prices (see Brown *et al.*, 1991). The reasons are that level and structure address somewhat different issues and that different types of regulation are available for each. The *price structure* predominantly deals with short-run allocation of existing plant and with the distribution of benefits between different customer groups. In contrast, *price level* regulation has a longer-run perspective and deals with distribution of risks and net benefits between customers overall and the regulated firm. The latter is known as rent extraction. Equally

[1] In this paper, we are totally neglecting environmental issues, implicitly assuming that they are taken care of through other policies that are reflected in a regulated firm's costs. In principle, we could have included environmental policy in output pricing. This would be appropriate if the output caused direct environmental harms.

important, price level regulation is decisive for the regulated firm's incentives for cost minimization.[2]

We begin with price level regulation. The options to be discussed are conventional rate-of-return regulation, price caps (based on indexed adjustments), yardstick regulation and profit sharing with consumers. We finish the section with remarks on giving regulated firms a choice from a menu of options.

2.1. Rate-of-return Regulation

Most U. S. electricity regulators currently practice rate-of-return regulation. Under rate-of-return regulation the firm's prescribed price level changes only if the firm's realized rate of return on capital deviates from an allowed rate of return. Thus, at least in theory, rate-of-return regulation is fully cost based. It fairly distributes cost reductions and cost increases between customers and the regulated firm, but would provide little incentives for cost reductions. Such incentives only come from "imperfections" in the practice of rate-of-return regulation. Two imperfections stand out. The first imperfection is that the allowed rate of return usually exceeds the cost of capital. As a result, regulated firms may have some interest in selling more output than an unconstrained monopolist but may also want to use inefficient input combinations (the Averch-Johnson effect). The second imperfection is that rate cases are costly and time consuming, resulting in a regulatory lag. During such a lag the firm can reap rewards from efficiency improvements and has to bear the burdens from inefficiency. Rate-of-return regulation will provide incentives for increased efficiency, for example, if the environment is stable (no inflation, no technical change, and no changes in interest rates). Under such circumstances regulatory lag will be long, giving the transmission company time to improve efficiency. The lag has, however, caused problems in adapting prices to changing environments, something that in the past came close to regulatory expropriation of electric utilities when input prices changed drastically (Joskow and MacAvoy, 1975).

[2] The distinction between price level regulation and price structure regulation is related to the Laffont and Tirole (1993) "incentive-pricing dichotomy". If this dichotomy holds, allocative pricing problems can be separated from cost-reducing incentives.

Because rate-of-return regulation is critically dependent on the rate base (the value of the firm's assets), the regulator has to scrutinize the firm's capacity investments. In the last few years U. S. regulators have questioned and successfully denied cost overruns or entire investments in the rate base. While an efficiency rationale can be given to this type of regulatory behavior (Gilbert and Newbery, 1988; Lyon, 1991) it does raise moral-hazard issues and increases regulatory uncertainty. As a result, theoretical models generating overinvestment under rate-of-return regulation may no longer be applicable (if they ever were). A potential underinvestment problem holds true in particular for investments with high cost risks to begin with, *e.g.*, nuclear power plants.

Summing up, rate-of-return regulation provides weak incentives for cost minimization. On the positive side, rate-of-return regulation in the U. S. has evolved from a fairness doctrine that has provided it with substantial commitment power.[3]

2.2. Price-cap Regulation

Price-cap regulation was invented as an alternative to rate-of-return regulation that would provide steep incentives for cost minimization by making regulated prices independent of the costs of the regulated firm. At the same time, risks from exogenous cost changes would be limited by an adjustment formula. Under price-cap regulation the firm's price level has to remain at or below a cap that moves over time, at an *exogenously* determined rate. The formula for this rate usually contains three distinct elements:

1) An adjustment factor for the economy's price level. This inflation adjustment can be seen as representing the firm's unspecified input prices or, more likely, the inflationary loss of consumers.

2) One or several adjustment factors for specific inputs or cost items that are passed through to consumers. These include some tax

[3] With all its drawbacks rate-of-return regulation, in the U. S., has to be considered a serious alternative to any form of incentive regulation. The reason is that (based, *inter alia*, on the 1944 Hope decision by the U. S. Supreme Court) regulated firms (and possibly customers) can always induce regulators or courts to revert to rate-of-return regulation if incentive regulation deviates substantially from rates of return that simply cover cost of capital. This is something where other countries would differ substantially from the United States.

and fuel adjustments. Besides for price changes of equipment, construction and labor, the major question for capital-intensive firms is "Should there be a price adjustment for economy-wide interest rate changes?" Such an interest rate adjustment would be entirely different from rate-of-return regulation because it would follow interest rate changes by simple formula (*e.g.*, linked to the prime rate). Such an adjustment makes sense even if part of the firm's debt carries fixed interest rates, because the firm's overall opportunity cost of capital is variable.

3) A general productivity adjustment factor, *X*. This would, ordinarily, reflect forward-looking productivity improvements, coming from technical change, economies of scale and adherence to cost minimization. Such improvements can be projected based on long-run trends from the past and from knowledge of the relevant technology. Contrary to telecommunications, the energy industries have not seen consistent long-run productivity trends that would lend support to positive levels of *X*. In a move from rate-of-return regulation to price caps it may, however, be appropriate to add a "consumer dividend" of, say, .5% to an *X* factor that is based on historic productivity data. The U. S. Federal Communications Commission (FCC) has, for example, done this, when introducing price-cap regulation for AT&T in 1989, arguing that price caps would provide incentives for productivity improvements that should be shared with customers.

The first two factors should be designed in such a way that they cannot be influenced by the firm but together correlate closely with the firm's overall input price level. If, in addition, the *X* factor provides a realistic assessment of the firm's productivity change over time, then the price cap will trace the firm's overall cost level. Although the price-cap formula adjusts for external cost and demand changes that a regulated firm may experience, cumulative deviations from normal profits could, over time, reach positive or negative magnitudes unacceptable to the regulator. This would happen because: *a)* not all cost factors are covered, *b)* external developments differ from firm-specific cost factors, and *c)* the *X* factor is at best a forecast for productivity changes. As a result, price cap formulas need to be revised every few years (or deregulation has to occur). These revisions tend to be based on the firm's achieved and expected rate of return. Hence, price-cap regulation is often viewed as similar to rate-of-return regulation, however, with a longer and pre-specified regulatory lag.

To summarize, the strong incentive effects of the price-cap formula for cost minimization are, in practice, limited by the necessity to contain distributionary effects and to prevent extreme rent transfers.

2.3. Yardstick Regulation

Under yardstick regulation the firm's price level is capped with reference to some yardstick. The yardstick can be an efficient cost level, such as long-run average incremental costs (the TSLRIC or TELRIC used by the FCC[4] — or benchmark costs used in Chile for electricity regulation) or the average level of costs achieved or of prices charged by comparable firms in the same industry.

The FCC approach of cost estimation is very tedious for network industries, in which investment costs depend very much on local circumstances. It is far too tedious for annual price adjustments but may be justified on a one-time basis, for starting prices.

Cost comparisons with other firms as the basis for yardstick regulation would allow the regulator to provide optimal incentives and leave the regulated firms no rents, to the extent that these firms face similar demand and cost functions and are subject to the same random shocks as the regulated firm. The problem is that regulated firms, even in narrow sectors, such as electricity transmission, can face vastly different demand and cost functions (for example, due to different terrain) and can be subject to idiosyncratic shocks. Yardstick regulation then loses its effectiveness in providing incentives and in limiting firm rents.[5] For example, a firm with very unfavorable cost conditions may be unable to finance investments if it can only charge prices based on firms with average conditions. Conversely, a firm with very favorable cost conditions may be able to reap excessive profits. Nevertheless, yardstick regulation may be the right approach, for example, if regulators have limited data sources. For example, in 1997, the European Commission recommended interconnection charges for tele-

[4] TSLRIC stands for total service long-run incremental cost and TELRIC for total element long-run incremental cost. See FCC (1996).

[5] Part of the cost differences between firms can be eliminated through econometric analysis. Econometric estimates can, for example, eliminate the effects of network density or of regional wage differences. There remain, however, many unexplained differences other than those relating to firm efficiency. This can create fairness issues if the firm is rewarded or penalized for such non-efficiency related differences.

phone companies based on the average interconnection charges of OECD countries with the lowest such charges. The reason for this procedure was that domestic cost data on interconnection were not readily available to national regulators in the European Union. At the same time, there was the presumption that efficient costs of interconnection would not differ substantially across industrialized countries. Thus, when using yardstick regulation regulators face a tradeoff between their own ability to measure benchmark costs and the potentially serious distributional issues that arise when other firms are used as yardsticks. If the own abilities for cost measurement are limited and if costs/prices of other firms are deemed similar to those of the regulated firm then those other firms can be used as a yardstick.

2.4. Profit-sharing (Sliding Scale) Regulation

Under price caps and yardstick regulation the regulated price level depends on variables outside the firm's influence. This provides strong incentives, but may provide outcomes that are perceived as unfair. Having consumers share the firm's profits addresses this fairness issue and, at the same time, provides at least some incentives for cost reduction. In this case, the firm's price level is adjusted by a specified profit share times the achieved rate of return on the firm's revenues. Since profits first have to be measured, the adjustment occurs with one period lag. Thus, if the firm's economic profit this period is π, its revenues are R and the sharing parameter is s, then the firm has to reduce its prices next period on average by a factor of $s\pi/R$.[6] The beauty of profit sharing is that, in an *ex-post* view, it treats the customers as shareholders of the company while, seen *ex ante*, it is a sharing of risk and an incentive device for the firm. For sufficiently short lag periods, the larger s, the smaller the incentive of the firm to reduce costs and the smaller the risk faced by the firm. With s approaching 1 the firm can keep any excess profits for only one period but also has to face losses only for one period. In this case profit sharing approaches cost-plus regulation without the 'plus'. With s vanishing, profit sharing approaches total deregulation of prices or pure price-cap regulation (with an infinite regulatory lag). For $0 < s < 1$ profit-sharing regulation has a number of interesting properties.

First, in a changing environment profit sharing, due to its lagged application, has somewhat erratic long-run dynamic effects on prices and financial performance. For example, a one-time cost reduction would lead to a simultaneous profit increase. Sharing this profit means that consumers receive a price reduction in the next period which, because costs are back to the old level, lead to a loss at that time. This loss then triggers a price increase next period, but not by the full amount. Profits would be converging from below to zero profits over time. Compare this to a permanent cost reduction, which will lead to a one-time profit increase, followed by a gradual price and profit reduction over time. If several cost and demand changes occur simultaneously or in short order there will be compound effects that can be hard to predict. These effects decline over time, with the speed of reduction depending on the profit share. However, there may be some risk of long-term losses to the firm. This specifically holds for inflation, which the firm may never catch up with. That is why customers may have to share 100% of all losses (after some lag period), or profit-sharing regulation may have to be combined with some adjustment formula for inflationary or input price changes.

Second, in terms of commitment power profit sharing may outperform other types of incentive regulation because of its built-in fairness and self-correction. The regulated firm is allowed to keep only part of its profits from windfall or superior efficiency; and consumers almost immediately share in the benefits. Consumers, therefore, can be happy about large profits because those trigger subsequent large price reductions. However, the lack of full loss sharing could induce regulated firms to revert to rate-of-return regulation.

Third, profit sharing can have similar incentive power to price caps, depending on the length of the regulatory period and on the sharing parameter. A major difference, however, is that, under price caps, there is no adjustment for profit changes before the review period and then there is likely to be a large or full adjustment while, under profit sharing, there is a partial adjustment in each period. Through the compounding effect of profit sharing the incentive for cost reductions is reduced by more than the share parameter (up to 100%, if the discount rate is zero and the cost reduction only lasts one period).

Fourth, if profit sharing is applied every period it has high

[6] Technically, profit sharing is often achieved by making the shares depend on achieved rates of return relative to target rates of return. The sharing parameter s may then itself depend on the amount of deviation from the target rate of return. In practice, the sharing is often set at 100% if rates of return are above and below certain threshold levels. Profit sharing is rarely done by direct payments to consumers because those raise fairness and moral hazard issues.

administrative costs because each time it resembles a rate-of-return rate case.

Thus, overall profit sharing has the advantage of popularity but carries major incentive and financial problems, unless losses are shared 100% by customers.

2.5. Menus

Our discussion of types of price level regulation has revealed that none of them is ideal, but that different mechanisms succeed or fail under different circumstances. The question is if hybrid schemes can make up for deficiencies by building on the strengths of individual mechanisms and avoiding their weaknesses.

What runs under the name of price-cap regulation, as it is practiced in several countries for electricity and in many U. S. states for telecommunications, is already such a hybrid in that it combines aspects of rate-of-return regulation, profit sharing, yardstick regulation and pure price caps.

One reason why single schemes are not ideal is that regulated firms hold private information that they can use to bend the effects of a particular scheme to their advantage. The regulator can make use of this tendency by offering firms menus consisting of combinations of schemes. The regulated firm will then select the scheme from the menu that is most adequate for it in that it maximizes expected profits among all schemes on the menu. The type of menu most commonly suggested consists of various blends between price caps and profit sharing. The firm could then choose between different X factors and profit shares, s, of consumers. The larger the X factor, the smaller the profit share of consumers would be. For example, $X = 0\%$ could be associated with $s = 100\%$, $X = 1\%$ with $s = 80\%$, and so on, until $X = 5\%$ and $s = 0\%$. The rationale is that a firm that expects to gain a lot in productivity would be more willing to commit to a high X. Thus, the most efficient type of firm would self select into the steepest incentive scheme (Laffont and Tirole, 1993).

In practice, menus are difficult to design and it is hard to set the right parameters. Also, customers tend to be unhappy about the choice from the menu made by the firm.

3. Price Structure Regulation

Disputes about the type of price-structure regulation arise with respect to consumer groups and a regulated firm's competitors. Consumer groups generally want the regulated firm to charge low prices for them relative to others, while the regulated firm's competitors want the firm to charge high prices in the markets where they compete. Under monopoly regulation there are no such competitors. However, for example in electricity, some generation and distribution companies or large industrial consumers are able to bypass the public transmission network. These potential bypassers may compete with other firms that cannot bypass the network. In this case they may want the transmission company to charge high prices so that they can bypass the network whereas their competitors cannot. If the price structure is heavily (cross-) subsidized incentives for entry or bypass may be created or enhanced, thus stranding investments by regulated firms.

The four types of regulating the monopolist's price level that we discussed above can be combined with several different types of regulating its price structure. We briefly consider three main types:

- Fully distributed cost pricing
- Price bands
- Flexible price structures

3.1. Fully Distributed Cost Pricing

Cost attribution formulas determine price structures by distributing the costs of the firm among its outputs and then making price structures depend on the costs thus allocated. This is usually known as fully distributed cost pricing. There are many ways to distribute not directly assignable costs among outputs, which makes this procedure very arbitrary (Braeutigam, 1980). However, it does have a long tradition in accounting and regulation and therefore is often the *status quo* against which new suggestions have to be measured. In particular, rate-of-return regulation has traditionally been associated with fully distributed cost pricing.

3.2. Price Bands

Bands with upper and lower price limits allow the firm some limited flexibility in changing its price structure while giving consumers assurances that they are protected from large price increases.

Numerically prespecified bands are routinely used in the U. S. for the regulation of access prices of the local telephone companies. In contrast to complete flexibility in the price structure, the regulator can commit more easily to prespecified bands. In the case of monopolists, lower limits are probably unnecessary because predatory pricing is of little concern. However, bands may provide assurances that specific customers are not favored by the monopolist.

Economically specified bands are defined by economically meaningful upper and lower limits on prices. Antitrust economists usually argue for stand-alone costs as an upper limit and incremental costs as a lower limit, based on cross subsidization and competition considerations. Prices above stand-alone costs subsidize others and could never be maintained indefinitely with free entry. Prices below incremental costs are subsidized by others and would never be maintained indefinitely by a profit-maximizing firm (and could not be maintained by a multiproduct firm in contestable markets). Although these upper and lower bounds are economically compelling and appear to have fairness acceptance, they suffer from measuring problems for both incremental and stand-alone costs. That is precisely why such bands appear to be more relevant for antitrust than regulation. Such bands would also be important for any prices that remain outside of regulation (*e.g.*, for optional prices discussed below) and for the regulation of integrated utilities that face competition on one level.

3.3. Flexible Price Structures

Complete flexibility in the price structure sounds like lack of regulation. It is, however, in practice severely limited by price-level regulation. If the firm has to stay on or below a regulated price level it is limited to price structures with which it at least breaks even. This choice of price structure obviously becomes more constrained the tighter the constraint on the price level. An unconstrained profit-maximizing monopoly firm will implement an efficient price structure, though at an inefficiently high level. By providing the right constraint on the firm's price level (in the form of a price index) regulation can benefit from the firm's natural tendency toward an optimal price structure. This is what price cap regulation is trying to make use of. However, fairness concerns or other than efficiency concerns may impose additional constraints on the firm's flexibility to choose its price structure.

The freedom of price structure can be restricted through the introduction of *baskets*. Each basket contains a subset of services for which a specific price level has to be maintained while, within the services of a basket, the price structure may be flexible. Thus, the price structure between baskets tends to be rigid. In the case of electricity transmission separate baskets could, for example, refer to native loads and third party loads, or they could refer to different zones.

Flexibility for *optional* prices allows the firm to offer consumers (nonlinear) price options in addition to regulated prices. Thus, there could exist a regulated price structure along with an unregulated optional price structure. Since each customer has the option always to buy at the regulated prices, customers would be protected. At the same time the firm can increase its sales and its customer base through attractive offerings. This could become important for a regulated firm faced with bypass.[7] While optional prices appear to be ideal for final (*i.e.*, residential) consumers, they may carry problems for commercial customers that compete with each other. For a commercial consumer it may actually be bad if an optional price is offered that is attractive for its rival but not for itself, because that might deteriorate its competitive position. Another problem of optional pricing can come from incentives for quality deterioration. The regulated firm may offer an optional tariff and, at the same time, deteriorate the quality under the regulated tariff. To avoid this, quality monitoring and guarantees may have to be part of the regulated tariff.

[7] Optional pricing also affects the firm's price level. If optional prices are kept outside the price-cap level the actual quantities traded at regulated prices may become irrelevant as weights. In contrast, if optional prices are included they will have a feedback effect on regulated prices.

4. Regulation of Electricity Transmission

4.1. Specifics of Electricity Transmission

The traditional electric utility setup has been one of vertically integrated monopolies that provide generation, transmission and distribution services. The modern view is that competition in generation is feasible and desirable and that distribution should occur in locally separated monopolies. In contrast, transmission should either be provided by independent regional transmission companies (TRANSCO's), or the grid should be leased to an independent system operator (ISO), who would be in charge of network coordination and generation dispatch. Since we want to concentrate on the monopoly regulation aspect, we consider the TRANSCO approach. This fairly neatly separates monopoly from competitive issues. After an extensive analysis of this case in the current section, we will, in Section 5, look at the case of a vertically integrated generation and transmission company that faces competition in generation.

What makes regulation of transmission networks particularly challenging is to set incentives in such a way that the transmission network ideally complements generation and distribution. This includes minimizing distances between power stations and demand centers for competitive alternatives, providing system reliability (frequency and voltage levels), smoothing load patterns, coordinating maintenance of power plants and providing emergency responses (Joskow and Schmalensee, 1983). All this has to be achieved for a commodity that is exceedingly hard to cost out.

TRANSCO's through their investment and pricing influence the amount of electricity transmitted through their network. For given investment their total costs are largely sunk, making capacity utilization their major short-run problem. Their major long-run problem is optimal investment, optimizing over the amount of expansion and minimizing costs of investment. Three issues need to be addressed in particular. First, transmission is a complicated service with severe externality problems that can at least be partially internalized by having a monopoly provider. Second, transmission networks exhibit economies of scale and lumpiness. Third, the main "variable costs" of transmission come in the form of power losses and congestion costs (which are opportunity costs). O&M expenditures vary little with usage (although inappropriate usage could lead to system breakdowns). Thus, the TRANSCO's variable costs are associated with hardly any expenses (unless it is responsible for making up power losses).

4.2. Suggested Price Level Regulation

Rate-of-return and yardstick regulation of TRANSCO's look quite unattractive. The cost minimization issues in transmission networks are probably too location specific to be handled by regulators as outside cost controllers and to provide yardsticks (on the basis of cost models or average costs of the industry). Thus, incentives from cost-monitoring under rate-of-return regulation and from industry-wide yardsticks are unlikely to be feasible in the case of a transmission company. However, rate-of-return regulation may have some crucial functions for initiating the price level and for long-run revisions.

For TRANSCO's we are thus left with price caps or profit sharing. We suggest a hybrid scheme, based predominantly on price caps.

The transmission price level crucially influences TRANSCO investment. First, in order to invest, the TRANSCO has to expect a rate of return on this investment covering the cost of capital. Second, transmission investment is risky because transmission links are lumpy and long lived and therefore the TRANSCO has to assess demand for transmission services over distant futures. This suggests that the price level has to contain either a buffer to accommodate risks or flexibility to adjust for risks *ex post* on short notice. A buffer can be built into the X-factor while flexibility in the short run comes from the cost adjustment and in the long run from rate reviews every few years.

Regulation of price levels would occur in three steps. First, initial prices need to be determined. Second, price levels will be adapted every period (if necessary and appropriate). Third, price levels will be reviewed every few years.

Regulation would begin with a determination of starting prices, usually the prices ruling before the introduction of the new regulatory regime. Such prices may not exist for transmission services, and, if they do, they are unlikely to be satisfactory starting prices for a TRANSCO. Thus, such prices need to be established with reference either to efficient prices (*e.g.*, level equal to long-run average incremental costs of expanding transmission capacity by a substantial lump) or to a target rate of return on historic costs. Selecting starting prices can be a complicated undertaking. The total embedded costs of a transmission system are easily determined. However, it is hard to translate these into a sensible starting price structure. One way to achieve a price structure is to distribute the embedded costs over kWh transmitted in a base period and over kW generation or load capacity on the grid. The problem, however, is to find correct percentages for the division between the two assignment methods. In principle, the costs assigned to kWh should reflect expected congestion costs and the costs assigned to kW the residual.

Subsequent periodic changes in price levels would, under our preferred approach, be ruled by a price-cap formula adjusting for input price changes (capital goods, labor and interest rates) and a productivity commitment *X*. For electricity transmission companies with few cost items that could be linked to the current level of inflation and with services that are not purchased directly by households needing protection from inflation, a general inflation adjustment might be inappropriate, although it is easily understood and implementable. As a result of computerization and increased electricity trade, some productivity improvements of transmission companies through better capacity utilization can be expected over time. However, it is hard to find empirical evidence in favor of a productivity-based *X* substantially above 0. The *X* factor should be on the order of 0-3%, reflecting a consumer dividend.

The price-cap review after, say, five years could use a rate-of-return approach, but should have both a backward-looking and a forward-looking component. For fairness reasons, some customer profit sharing for the past five years could come in if no consumer dividend was built into the *X* factor. This profit sharing could take the form of a limited one-time rebate, a one-time adjustment in the price level or an adjustment in the *X* factor. For efficiency reasons, the new *X* factor should otherwise be based mainly on the future possibilities of the firm.

4.3. Suggested Price-structure Regulation

4.3.1. Two-part Tariffs

Transmission pricing is the perfect case for the following conundrum. In the short run, the goal is optimal capacity utilization, something that can be done under sophisticated pricing that involves a large

amount of (partially private) information and reflects the cost of congestion to users. Because of fluctuating and inelastic demand, such pricing may lead to revenues that are unrelated to the capital cost of capacity. In particular, severe congestion can be highly profitable for the TRANSCO. Thus, the TRANSCO may have too little incentive to invest when new capacity is most needed; and the incentives for optimal capacity utilization and optimal investment are hard to coordinate. Optimal investment would mean that investment occurs at the margin when the marginal cost per unit of new capacity equals the expected congestion cost arising from not adding that unit.

How can the tension between optimal capacity utilization and optimal investment be overcome in a price regulation scheme? We are looking for regulated prices that are flexible enough in their structure to allow for efficient capacity utilization, yet will generate stable enough revenues to support capacity costs. They should also provide incentives to invest precisely when capacity utilization is getting too high. To achieve this, we suggest a *two-part pricing* to the price-cap index.[8]

Under this approach, the Transco could charge an average tariff that would follow the price-cap level by declining, for example, *X*% per year in real terms. The average would be defined as a price index with quantity weights.[9] The index would be defined on a set of fixed fees and variable fees. The basic idea is that optimal capacity utilization can be reached through the variable fee, while the incentive to invest can be captured in a combination of revenues from the variable tariff plus fixed fees. The background is that capital costs of capacity are steady while congestion costs are highly erratic. Thus, total revenues need to be fairly steady and reflect the influence of capital costs, while the part of revenues coming from current operation should reflect the ups and downs in capacity utilization. This can be done by creating a tradeoff between utilization tariffs and fixed tariffs. To achieve this, the price-cap index simply has to be a price index defined over two-part tariffs. In the simplest case, the quantity sold under a fixed fee is the number of customers buying under this fee.

[8] The formula can be applied just as well to other forms of price-level regulation, provided the price level is high enough to support investment.

[9] Such an index would have the form $PI = {}_i\Sigma p_i^t q_i^w / {}_i\Sigma p^{t-1}\, {}_i q_i^w$. If it is a chained Laspeyres index with previous period's quantities as weights it can also be written as $PI = {}_i\Sigma(p_i^t/p_i^{t-1})(p_i^w q_i^w / {}_i\Sigma p_i^w q_i^{w)}$.

Basically, whenever the TRANSCO increases its utilization tariffs as a result of congestion, it has to reduce the fixed fee accordingly. Incentives for efficient behavior are induced by using quantities for transmission services as weights for the variable fees so that total revenues (and resulting profits) will be higher the more efficient are capacity utilization and capacity expansion (*i.e.*, capacity utilization in the short and in the long run).

4.3.2. Variable Fees

The variable fees take care of the short term pricing issues of the TRANSCO. In particular, they could cover congestion costs, power losses and ancillary services. The variable fees could take simple or sophisticated forms. In that sense, the suggested regulatory approach is compatible with any desired transmission pricing approach of regulators and firms. Under a simple approach, the TRANSCO would set variable fees *ex ante*. For example, there could be peak and off-peak rates for different zones. Once the TRANSCO sets those rates, they would determine the level of fixed fees (following the price-cap formula). Within a given price-cap level the price structure could then be changed on short notice. For example, if fixed fees are paid on a monthly basis, the TRANSCO could also change its fee schedule monthly with, say, 15 days notice. This way, regulation could accommodate changes in demand and supply conditions.

A sophisticated approach would be spot pricing (Schweppe *et al.*, 1988). In this case, the variable fees might change almost instantaneously and could differ geographically, for example, by nodes or zones. At the variable fees it charges, the TRANSCO would then have an obligation to serve. So, there would be no non-price rationing. In this scenario fixed fees could not be set *ex ante*. Because of congestion, spot prices would fluctuate and, by definition, be unknown *ex ante*. The fixed fee would therefore have to be determined at the end of the period. These within-period adjustments, which make the TRANSCO fulfill the price-cap constraint, can be interpreted as *premia* and *penalties*. If there is too much congestion in a period the variable fee will automatically adjust upwards (to equate supply and demand) and therefore the TRANSCO will have to reduce fixed fees (across the board). Thus, the TRANSCO would be penalized for congestion. If there is less than expected congestion the variable fees will be lower than expected and hence fixed fees will be increased, generating a premium for the TRANSCO. Because prices stay within averages, the firm can make extra money only through expanding quantities (as long as the average incremental costs of expansion are below the price level).

In principle, the same incentives with premia and penalties would hold for the case of simple variable fees that are set *ex ante*. The TRANSCO would set those in such a way that the existing transmission capacity is optimally utilized. Foregone loads and load shedding mean reduced revenues. Pricing responses to avoid these reductions go in the same direction as under spot pricing (only less responsive) and the compensating changes in fixed fees are similar.

4.3.3. Fixed fees

Fixed fees fulfill several functions. In particular, they help pay for the difference between total costs and variable fees. Although variable fees could cover total costs, it would be pure coincidence if they did so efficiently. Prudent capacity reserve margins, for example, could force efficient variable prices far below cost-covering levels. Fixed fees could also pay for backup transmission capacity for customers ordinarily bypassing the grid of a TRANSCO. Pérez-Arriaga *et al.* (1996) contend that variable fees at marginal costs would pay substantially less than total costs.

Fixed fees will have to be discriminatory in the sense that different customers pay different amounts of fixed fees.[10] Otherwise, large customers would be heavily favored. Fixed fees should fulfill at least two requirements. They should be fair (subsidy-free) and they should not depend directly on individual usage. If they depended on usage they would not be fixed but rather variable. What they should depend on is the transmission *capacity cost* caused by a customer and/or the customer's *net benefit* derived from the network's use. The range between capacity costs caused by the customer and customer benefit (both net of variable contributions) would define fixed fees that are free of (cross-) subsidies. A proxy within this range could be total generating capacity or load of a user connected to the grid (the size of

[10] However, they need not to be discriminatory in the sense that they result in different average prices paid by different customers.

the fuse).[11] More efficient (less distortionary) but harder to determine would be fixed fees proportional to net benefits.

4.3.4. Weights of the Price-cap Index

Besides the immediate penalties/premia created by the within-period adjustments forced by the price-cap constraint there is a between-period (annual) adjustment due to the dynamic nature of the formula. This between-period adjustment depends on the precise formula used for determining weights of the price-cap index. As shown below for a stylized example, the more (relative) weights deviate from optimal quantities the more will the outcome deviate from the optimal outcome. Thus, it is important to set price structures approximately right in the initial period so that convergence to optimal prices does not take too long.

The most important weights in use or discussed in the literature are:

1) Quantities of the previous period (chained Laspeyres price index)
2) Quantities of the current period (Paasche weights)
3) Weights that are fixed over time (fixed Laspeyres weights)
4) Projected quantities (Laffont-Tirole weights)
5) Flexible weights (resulting in an average revenue constraint)

1) The most common formula in actual price-cap regulation uses quantities of the previous period as weights (chained Laspeyres price index). These weights have substantial advantages. They are easily verifiable and usually close to the optimal weights. Also, the change in profits resulting from these weights is usually smaller than the change in welfare. Their drawback is that, in times of major demand and cost changes, they can differ substantially from optimal weights, resulting in suboptimal prices and output levels. However, once costs and demand stabilize, these weights converge to optimal weights.

2) Paasche weights, which use current quantities, usually have the problem that quantities are not known before the end of the period, meaning that allowed prices would also be known at that point in time only. However, if variable prices simply reflect congestion they would be determined by the market and the resulting weights would be used only to determine fixed fees. Even if this solves feasibility problems, Paasche weights do not have very desirable properties. In particular, there is no intrinsic tendency to converge to optimal weights over time. Also, the change in profits resulting from these weights is usually larger than the change in welfare. This may have to be countered by a larger *X* factor.

3) While fixed weights are not subject to manipulation, their problem is that they do not adjust to changing circumstances and are therefore often far off from optimal weights.

4) Predicted quantities as weights are optimal to the extent that the predictions are correct. However, if demanded quantities expand, the change in profits resulting from these weights is usually larger than the change in welfare. This may require a larger *X* factor to compensate. Contrary to the other weights discussed so far, determining these weights requires sophisticated analysis by the regulator. We discuss them below in connection with global price caps (Section 5.3.2).

5) Average revenue constraints have the advantage of extreme simplicity. However, they contain a certain arbitrariness (that they share with fully distributed cost pricing). In order to define an average revenue total revenue has to be divided by a single type of quantity units. For example, for an electricity generation or transmission company the most obvious unit would be kWh, masking differences in voltage levels, reliability, time of day, location, etc. Because of this arbitrariness it is hard to know how far the resulting weights differ from optimal weights. As shown by Sappington and Sibley (1992), average revenue constraints could create incentives to introduce inefficient forms of two-part tariffs.

Although previous quantities are not optimal, they probably strike a balance between practicality and efficiency. Below, we try to show that lagged weights provide good operating and investment incentives. The alternative would be to set some fixed benchmark weights, based on ideal capacity utilization.

[11] One has to be careful, though, not to create perverse incentives. For example, if such a fixed fee would prevent a generating company from building a power plant, then this may actually increase transmission congestion and may increase the required transmission capacity. Thus, additional generation capacity may appear to contribute negatively to transmission costs. In this sense, generation and transmission can be substitutes. The important issue is to minimize the sum of transmission, generation and distribution costs for given electricity consumption.

4.4. A Simple Model

4.4.1. The Price-cap Constraint

We will now formalize the price-cap constraint and derive some properties through simple profit maximization.

Assume that the firm faces a price-cap constraint based on the average of prices it charges. In this case, the average is defined by a price index over two-part tariffs. The subscripts of fixed fees (the *F*'s) run over all potential consumers $j, j = 1, \dots, N^t$, reflecting discrimination in fixed fees. The subscripts of usage fees (the *p*'s) can run over sub-periods, nodes, voltage levels, etc., all covered by $i, i = 1, \dots, M$. The quantities (the *q*'s) corresponding to usage fees could be kWh (or some other unit for ancillary services). Summation is over *i*'s and *j*'s. Superscripts refer to periods (years). We neglect price uncertainty with an *ex post* determination of fixed fees.[12]

The price-cap constraint verbally described in the previous section would, in period *t*, have the form:

$$_i\Sigma p_i^t q_i^w + {}_j\Sigma F_j^t\, \delta_j^w \leq ({}_i\Sigma p^{t-1}{}_i q_i^w + {}_j\Sigma F^{t-1}{}_j\delta_j^w)\,(1-X)$$

or

$$({}_i\Sigma p_i^t q_i^w + {}_j\Sigma F_j^t\delta_j^w)/({}_i\Sigma p^{t-1}{}_i q_i^w + {}_j\Sigma F^{t-1}{}_j\delta_j^w) \leq (1-X). \qquad (1)$$

Here, δ_j^w is either '0' or '1', depending on whether a particular firm *j* is a subscriber at the time. In more compact vector presentation we can write the price cap as

$$(\mathbf{p}^t\mathbf{q}^w + \mathbf{F}^t\delta^w)/(\mathbf{p}^{t-1}\mathbf{q}^w + \mathbf{F}^{t-1}\,\delta^w) \leq (1-X), \qquad (1a)$$

with **p** an 1x*M* vector, **q** an *M*x1, **F** an 1x*N*, and δ an *N*x1 vector. For the moment, we neglect '*X*' and set it to zero. This brings out more clearly the basic tradeoffs involved. Also, assume that tariffs are such that within realistic pricing options the number of customers is fixed.

[12] Such *ex post* adjustment could involve interest charges, strategizing on discrimination and the effects that differences between before/after prices could have on the number of customers (generators, distribution companies and final users).

In the simplest case of a given number of users, *N*, and only one usage charge and one fixed fee, equations (1) and (1a) can be reduced to:

$$F^t \leq F^{t-1} + (p^{t-1} - p^t)q^w/N\,. \qquad (2)$$

or:

$$\Delta F \leq -(\Delta p)q^w/N, \qquad (2a)$$

where Δ signifies change.

(2a) shows nicely how fixed and variable fees can be traded off against each other. The trade off factor is the ratio between the output weight and the number of customers.

Now, for the simple case, consider the firm's objective function in the short run:

$$\max \pi^t = p^t q^t + F^t N - C\,(q^t, K^t) \text{ s.t. (2) and s.t. } q^t \leq K^t. \qquad (3)$$

We assume a simple cost function $C\,(q^t, K^t) = C\,(q^t, K^{t-1})$ for $q^t \leq K^{t-1}$ and $C\,(q^t, K^t) = C\,(q^t, K^{t-1}) + I^t$ for $q^t > K^{t-1}$, with $I^t = q^t - K^{t-1}$. This reflects the long-run and sunk nature of the transmission grid.[13] Thus, costs only change through the addition of new capacity.

The first-order condition of (3) with respect to p^t under binding constraints (with μ^t as the Lagrange multiplier of the capacity constraint) is:

$$(\partial q^t/\partial p^t)\,(p^t + \mu^t - \partial C/\partial q^t) = q^w - q^t. \qquad (4)$$

At optimal investment we have $\mu^t = 0$, which implies:

$$(p^t - \partial C/\partial q^t)/p^t = -(q^w/q^t - 1)/\varepsilon, \qquad (4a),$$

where ε is the demand elasticity.

Some results can be shown fairly easily, assuming the absence of strategic behavior by the TRANSCO (or by large customers) to influence the price-cap formula:

[13] $C(q^t, K^{t-1}) = 0$ would reflect the sunk nature of the grid. Our formulation includes this case as a possibility.

1) In times of excess capacity and stationary demand functions, usage fees will decrease, usage will increase and there will be no investment. The reason is that a decrease in usage charges leads to an increase in fixed fees that together generate an increase in total revenues. Since, with excess capacity, there is no cost increase for more usage, net profits must increase.

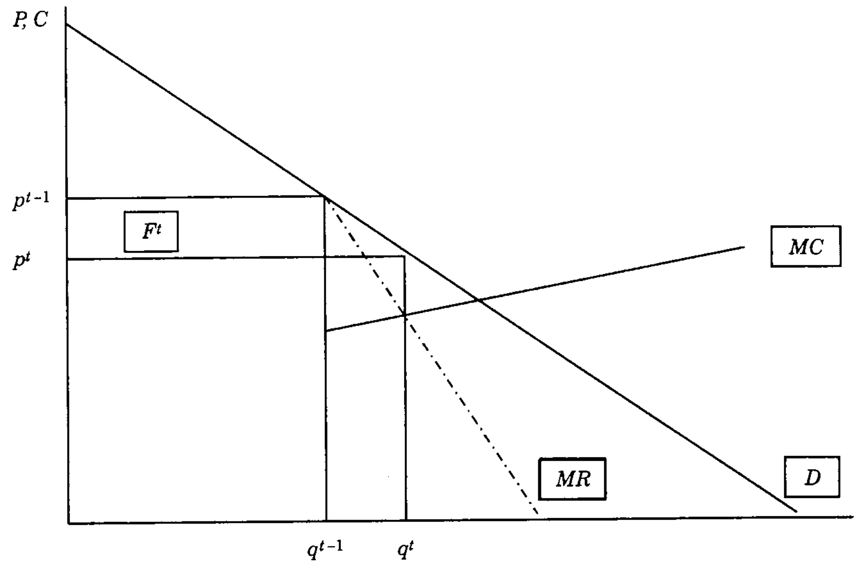
2) In times of binding capacity constraints the usage charge equals the users' marginal willingness-to-pay. This says that we have ruled out non-price rationing. Thus, in this case, the usage charge is a pure congestion charge.

3) If congestion charges on the margin are higher than the marginal costs of adding capacity the firm will have an incentive to add capacity.

The strength of the investment incentive (and of the incentive to lower usage prices) depends on the weights of the price index used. If the price-cap is a Laspeyres index (with last period's quantities as weights) the incentives are such that the TRANSCO will not invest the full difference between the *status quo* capacity and the optimal capacity. The reason is that the firm faces a trade off between making extra money on usage (congestion) and the amount it can make by increasing fixed fees. The latter is restricted by the weights on usage and fixed fees. For a given number of customers the firm faces a "residual demand curve" similar to the case of a Cournot oligopolist. In Figure 1, this is the demand curve D starting from quantity q^{t-1}. This is because a change in the variable fee translates into a change in the fixed fee that, applied to last period's quantity, exactly equals the price change. Thus, any profit change resulting from a change in the variable fee applies only to the change in quantity, starting from q^{t-1}. As a result, the firm will behave like a monopolist on the "residual demand curve" (except that its marginal cost curve starts at zero quantity). Thus, the firm will want to invest if customer willingness to pay exceeds marginal cost of investment but will not invest to the point where the two are equal. Rather, investment will proceed over time and (with stationary cost and demand functions) converge to the optimum capacity. This is a conjecture based on earlier work of mine on this type of scheme (Vogelsang, 1989. See also Bertoletti and Poletti, 1997).

Note that (2) with last period's quantities as weights describes a Slutsky-type approximation to the total consumer surplus change. The mechanism is therefore an approximation to the Loeb-Magat scheme (Loeb and Magat, 1979). Only if weights correspond to the quantities traded in the optimal state will investment be optimal as well. This conjecture follows from the Laffont and Tirole price cap approach (Laffont and Tirole, 1993). It obviously holds for the case of myopic profit maximization. To see this, consider the first-order condition (4). In contrast, the corresponding condition for social surplus maximization would have zero on the right hand side. Now assume weights equal to the socially optimal quantity, q^*, at which customer willingness-to-pay equals marginal cost. Figure 2 depicts this case. The firm's "residual demand curve" would now start at q^*, and there would be no incentive to move to another quantity.

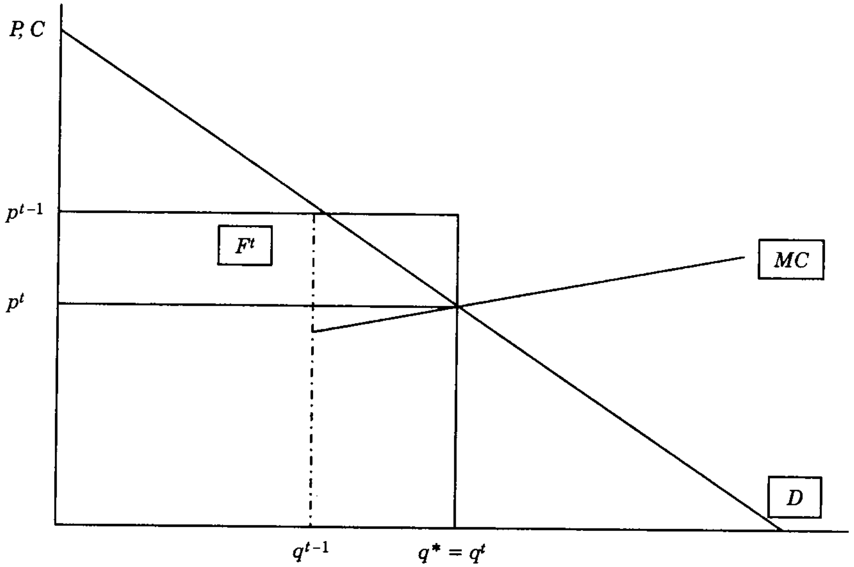
Figure 1. Price Cap with Laspeyres Index



The problem of using optimal weights is how to determine them without going through the motions of sophisticated regulation. Since, as can be seen from Figure 2, using optimal weights provides an opportunity for the regulated firm to earn at least the increase in social surplus, it should be possible to design a mechanism to make the firm reveal the optimal weights. Analogous mechanisms with a similar objective include the ISS-R by Sibley (1989).[14] Under this mechanism,

[14] For a slightly different mechanism in the same spirit, see Vogelsang (1990).

Figure 2. Price-Cap with Optimal Weights



the regulated firm would be able to offer an optional two-part tariff along with last period's tariff. Thus, customers can always choose last period's tariff. The optional tariff must therefore make them at least as well off. Thus, the firm maximizes profit w.r.t. the optional two-part tariff under the constraint that consumers get a specified surplus. The optimal variable fee then equals marginal cost and the fixed fee takes away the consumer surplus increase.

As discussed above, fixed fees proportional to transmission capacity costs caused by a user could be approximately efficient. If we find such a proportionality factor (for example, generation capacity connected to the grid) we can use a capacity charge instead of the fixed fee and use the proportionality factor as the quantity of capacity "consumed". This proportionality factor would then be the basis for quantity weights of what we have termed the fixed fee (instead of the δ's). If the proportionality factor is exact the firm would have strong incentives to expand capacity optimally.

We have, so far, only considered myopic profit maximization of the firm. In a fully dynamic setting with maximization of a discounted stream of future profits and with changing cost and demand functions things become more complicated and less predictable. If capacity expansion is required we can safely conjecture that capacity will converge to an optimal level as long as cost and demand functions do not shift over time (Vogelsang, 1989). This would hold only for expansion, since we had assumed that capacity cannot contract. We also know that changing demand and cost functions can cause problems, because the speed of adaptation of price-cap mechanisms is restricted by the speed of change in weights.[15]

4.4.2. The Effect of *X* on the Investment Incentive

Provided that the input price adjustment factors correctly trace the TRANSCO's input price developments the initial price level and the *X* factor jointly determine the TRANSCO's price level relative to average (and marginal) costs. If the firm's initial price level does not cover the average incremental cost of expansion there will, at least initially, be little incentive to invest.[16] Thus, setting the initial price level below average incremental cost of expansion would conflict with investment incentives and would therefore work well only if no immediate investment is required. Since we want to bring out investment incentives, we assume that expansion through investment would be optimal and that the initial price level exceeds the average incremental costs of expansion. In order to preserve incentives to invest, *X* can, in the long run, not exceed the firm's productivity growth on new investment. Such productivity growth would come from (a) more efficient operation due to better incentives, (b) technical progress, and (c) economies of scale effects.

The precise investment effects depend on the forward-looking difference between price level and average incremental costs of expansion. This difference is captured largely in *X*. Increasing the difference (decreasing *X*) has two effects (assuming that the difference is positive and below the unconstrained monopoly level). First, a larger difference makes investments more profitable, leading to increased investment incentives. Second, a larger difference means a higher price, which

[15] For demand changes, see Neu (1993); for cost changes, see Fraser (1995). Sappington and Sibley (1992) show that there may be a strategic incentive to use (inefficient) two-part tariffs if the price-cap is in the form of an average revenue constraint. This differs from our case where the two-part tariffs are part of the price-cap.

[16] The regulator could force such investment by imposing penalties for deteriorating quality of service outside the price-cap constraint but that would be outside the approach taken in this paper.

decreases the demanded quantity. Under linear pricing and under pure price rationing, the second effect dominates. This means that, without non-price rationing, investment would increase in X as long as the new price level stays below average incremental cost of expansion (Cabral and Riordan, 1989). Under our two-part tariff scheme this does not necessarily hold. As long as the number of customers is not affected by the fixed fee, the same variable fee stays optimal for the firm. We can see this by adapting (2) to the case of $X \neq 0$. Then (2) becomes:

$$F^t \leq (1-X)F^{t-1} + [(1-X)p^{t-1} - p^t]q^w/N. \qquad (2')$$

As can be seen, a change of X in this constraint does not affect the first-order condition (4). The investment incentive in the short run would therefore be unaffected by changes in X, as long as investment remains profitable. In contrast, it appears that long-run dynamic investment incentives could be affected.

4.4.3. Market Power of Generators as Users of Transmission Lines

The market for transmission services may be dominated by a single customer. This would, for example, hold if the TRANSCO was originally split off from a vertically integrated electric utility, whose set of generation plants is unaffected by the vertical separation. A dominant customer could try to exercise monopsony power and/or raise rivals' costs. As indicated above, raising rivals' costs can be addressed by price-cap baskets and by rules governing optional pricing. Monopsony power expresses itself in less purchases, in order to get lower prices. This could only work to the extent that a reduction in purchases lowers the dominant customer's average transmission prices.[17] That, however, would be largely precluded by the two-part tariff price-cap scheme. On the contrary, if the TRANSCO mimics the price-cap constraint in the tariff for the dominant purchaser average price paid would increase, as purchases are reduced.

However, for two reasons large generators may actually want to reduce transmissions. The first reason is to reduce transmission prices over the long run by reducing the usage weights and usage prices. The second reason is to keep the transmission network small, thereby reducing competition in generation. The first reason is unlikely to work, because average price is likely to increase in spite of lower weights on usage. The second reason may require collusion to be effective, because otherwise small generators could free-ride on this strategy. It could also be counter-acted by making distribution companies and industrial customers buy transmission services unbundled from generation.

[17] We are here neglecting the possibility that the dominant purchaser may have market power in the market for electricity generation.

4.5. Conclusion on Transmission Prices

According to Green (1997) 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.

We have concentrated our discussion on the first two of these principles, which our price-cap scheme with two-part tariffs should fulfill. Included in these two principles is the quality of service. 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.

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 that would signal transmission investment costs incurred by customers.[18] 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. Contracting could also be used by customers as the basis for becoming resellers.

Principle 4 (cost recovery) can be achieved through initial rates that reflect embedded investment costs. In this case, the X factor 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 should be positive, forcing the firm to reduce its costs through investment. Vice versa, if embedded average costs are below forward-looking incremental costs, X should be negative.

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 trade off 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 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.

[18] It is actually not clear that individual customers need stable transmission prices. Variable spot prices, for example, are likely to reflect variable conditions of customers and may actually reduce swings in customers' earnings.

To the six principles for transmission pricing proposed in Green (1997) a seventh, regulatory, principle should be added. It is that regulation should not stand in the way of innovative pricing by market participants. This means that the regulatory price mechanism should be flexible enough to accommodate both simple and sophisticated transmission tariffs. This is something the above mechanism was designed to do.

5. Vertical Integration of Generation and Transmission

5.1. The Efficient Component Pricing Rule (ECPR)

Regulation of transmission pricing differs between a TRANSCO and an integrated generation and transmission company mainly because other electricity generators using the transmission grid compete with the integrated firm. Thus, the transmission grid is an essential facility (or bottleneck input) supplied to these independent generators by a competitor (who may also be the dominant generator).

Recent years have seen substantial dispute among economists about regulated pricing of bottleneck inputs sold to competitors. The pricing rule most hotly discussed in the literature is known as the efficient component pricing rule (ECPR).[19] It says the integrated company should charge a transmission price equal to the incremental resource costs of transmission plus the so-called "opportunity cost" of transmission. This opportunity cost is the foregone profit contribution of the integrated company by providing transmission to a competitor who might use transmission to displace generation services provided by the integrated company. Thus, the ECPR is driven by the integrated company's wholesale electricity prices. If (a) transmission and wholesale electricity are generated in fixed proportions and if (b) the integrated company's and the independent generators' wholesale electricity are perfect substitutes, and if (c) generators take the integrated company's price of the competing wholesale electricity as given, then the opportunity cost is simply the profit contribution or quasi-rent

[19] The ECPR is widely attributed to Willig (1979) and Baumol (1983). For an extensive discussion, see Baumol and Sidak, 1994 and the Winter 1994 edition of *The Yale Journal on Regulation* and in the Fall 1995 issue of the *Antitrust Bulletin*.

generated by the integrated company's wholesale electricity (simple ECPR). Otherwise, the opportunity cost may be a fairly complicated term, reflecting cross-elasticities of wholesale electricity between different vendors, technical substitution and types of competition (sophisticated ECPR). The sophisticated version would apply to the relationship between electricity generation and transmission where, for example, (a) does not hold.

The main peculiarity in approach taken by the proponents of the ECPR is the assumption that the price for the wholesale electricity would be given (and chosen optimally) and that the only function of competitive entry is to provide part of generation at lower cost than the integrated company. The ECPR is therefore a partial rule that deals only with a specific aspect of electricity pricing and competition. It has nevertheless proven to be highly policy relevant. The reasons are that, with the simple version of opportunity cost:

- It is easily understood and practiced,
- It is often embraced by incumbents,
- It does not require a change in (regulated) prices of final services and does not interfere with politically popular cross subsidies.

With the more sophisticated version of opportunity costs the ECPR is also theoretically quite attractive but much more demanding on the regulator (see Armstrong, Doyle and Vickers, 1996).

The simple version of the ECPR is formally very similar to the requirement of imputation. Imputation means that the integrated firm may not price transmission at a lower price to itself than to others. Imputation is imposed in order to eliminate foreclosure incentives of the integrated company's simultaneous pricing in the transmission market and in the wholesale electricity markets. Because internal prices, in contrast to external transaction prices, do not usually have direct allocative effects (because internal payments cancel each other out), they can be used only as an accounting device to discover cross-subsidies. The imputation requirement shall thus guarantee that the wholesale electricity stage is not cross-subsidized. Laffont and Tirole (1996) equate the imputation requirement with the ECPR. However, imputation implies upper bounds for transmission charges (or minimal internal transfer p·ices), while the ECPR declares these upper bounds to be optimal.

5.2. Ramsey Prices

Theoretically optimal transmission prices can be determined under the Ramsey pricing approach taken by Laffont and Tirole (1993 and 1994). This approach simultaneously determines optimal transmission and electricity prices, and it makes no *a priori* assumptions about demand relationships, technology and type of competition. Rather, the assumptions vary, like in oligopoly models in general. Depending on which assumptions are made, the approach leads to different results. In general, these results are complex in that they have to deal with the integrated company's budget constraint, demand relationships, cost relationships and types of competition. This complexity reflects complicated relationships that need to be dealt with and is the price to be paid for general rather than partial optimization.

For example, a model by Masmoudi and Prothais (1994) on telecommunications would yield transmission prices with the following components:

- The marginal cost of transmission.
- A Ramsey markup including market shares of generators and the type of competitive interaction between them.
- A differential efficiency term reflecting the difference in efficiency between the integrated generator and the independent generators. This term has two opposing components: The more efficient an independent generator, the more it should produce relative to the incumbent, thus the lower the transmission price. Conversely, the more stringent the integrated firm's budget constraint, the less weight is given to the entrant's efficiency.
- A transmission charge elasticity term relating the transmission charge to the entrant's electricity output. The less elastic this output is to the transmission charge the higher the transmission charge should be.

Also, the optimal electricity prices themselves would obey a complicated markup formula. In reality, a regulator cannot hope to capture all these effects at the same time.[20]

[20] These effects do not yet include incentive effects as discussed in Laffont and Tirole (1993). The absence of incentive effects can be justified if the incentive-pricing dichotomy holds.

5.3. *Price-Caps*

5.3.1. Price-Cap Options

Information about the relevant technological and demand properties necessary to derive Ramsey prices or the ECPR is either unavailable or squarely rests with the integrated company. This makes it virtually impossible to directly implement the Ramsey approach and the ECPR. One could argue that asymmetric or lacking information can be captured in Bayesian incentive schemes using subjective probability distributions. However, such schemes lose much of their power if uncertainty parameters are too vague and too many. Thus, we take the same price-cap approach to the integrated firm as we did to the TRANSCO. However, in addition we have to take care of the different markets for transmission and wholesale electricity and of the relationship between the two. This gives us four options to consider:

1) Global price-caps that include transmission pricing and electricity pricing in a single approach (the same basket).

2) Separate price-caps (baskets) for the transmission and generation markets.

3) Price-caps for generation, but no price-caps for transmission.

4) Price-caps for transmission, but no price-caps for generation.

5.3.2. Global Price-Caps

Laffont and Tirole (1994 and 1996) have made a strong case for global price-caps. They argue that making the integrated firm choose its overall price structure under a common constraint on the price level can align the incentive for optimal pricing in both markets. They do, however, assume that the price-cap index uses optimal weights to begin with. In addition, they want to reduce any incentives for anticompetitive behavior by imposing an imputation rule for transmission pricing in addition to the price-caps. Thus, any individual transmission prices would have to obey both the price-cap and the imputation rules.

Under global price-caps a condition like (4) above holds. Thus, optimal weights would be the correctly predicted output levels. Making such predictions looks doable for a regulator. However, it actually means solving the Ramsey pricing problem discussed above. This would be very hard and would make the use of price-caps superfluous because, by solving the problem, the regulator would know the Ramsey prices and therefore could prescribe them directly. Thus, in applying global price-caps one will probably have to compromise on weights that are either quantities of past periods or quantities projected from past trends. The *X* factor would have to be derived from a weighted average of productivity increases for transmission and generation.

In theory, global price-caps provide the integrated firm with the ability and incentive to generate Ramsey prices overall. The imputation requirement may reduce this ability but that would only be in those cases where Ramsey prices would imply market foreclosure of rivals. Nevertheless, global price-caps may be too bold for a regulator to implement. One reason is that it is common knowledge that the regulator cannot commit to a specific regulatory scheme in the long run. Thus, under global price-caps, the integrated firm may use aggressive tactics against rivals, in order to keep its overall market position, in case regulation changes in the future.

5.3.3. Separate Price-Caps

From the regulator's perspective, separate price-caps are less daring and therefore more acceptable than global price-caps. There would be a transmission price-cap like the one described in Section 4 above and a wholesale electricity price-cap. The latter would be restricted to the integrated firm, while independent generators would be free to set their prices.[21] The wholesale price cap could be quite similar in spirit to the transmission price cap. A two-part tariff scheme here would be equally compelling. Since transmission services are inputs for wholesale electricity, there should be a passthrough of transmission charges. This could also be used for imputation purposes.

Having separate price-caps for wholesale electricity and transmission is theoretically non-optimal because the separation restricts the integrated firm's freedom to rebalance prices between transmission and generation. However, just like separate price-cap baskets they provide the regulator with additional controls.

[21] In my view, regulation should be asymmetric, as long as the integrated firm is clearly dominating the market. Once that is no longer the case, one should deregulate the integrated firm rather than regulate the independent generators.

5.3.4. Price-Caps for Generation Only

Since electricity generation is more competitive than transmission, it may seem awkward to regulate generation but not transmission of the dominant integrated firm. One could here use the argument of proponents of the ECPR that voluntary negotiations between the integrated firm and the independent generator would automatically lead to the ECPR.[22] Thus, if wholesale electricity rates were regulated optimally, the ECPR would result in transmission charges that could be optimal. There are two problems with this. The first is that optimal regulation of generation is not guaranteed. In particular, the regulator usually cannot commit to a regulatory scheme that would leave profits of the integrated firm intact after it lost a lot of market share in generation.

Second, by keeping transmission capacity small (by charging high transmission fees) the integrated firm can influence the type of competition in electricity generation and the investments of competing generators. Thus, by deviating from the ECPR, the integrated firm could try to prevent competition in generation from happening.

Regulating generation but not regulating transmission seems to have it the wrong way round. This holds, in particular, because transmission is the key to competition in the electricity sector.

5.3.5. Price-Caps for Transmission Only

While generation will not be perfectly competitive, entry into generation has become easier and economies of scale are less pronounced than in the past. In contrast, transmission is likely to keep monopoly advantages for some time. As a consequence, transmission and wholesale electricity can be sufficiently separated so that (for some time) transmission can be regulated while the wholesale electricity services produced with transmission as an input can be left to market competition.

Price-caps for transmission would be similar to those described in Section 4 above. In addition, there would be the need for an imputation rule for the integrated firm not to sell transmission to itself at a lower price than to independent generators.

[22] This is not strictly true. In a bilateral negotiation over transmission services between the integrated firm and an independent generator the ECPR would be the integrated firm's threat point, while the independent generator's threat point would be not to operate. If there is any room for negotiation the integrated firm will deviate from the ECPR.

I conjecture that under any separate regulation of transmission pricing the wholesale electricity prices resulting under competition will *ex post* yield the ECPR (in the sense that the transmission charge will equal the incremental cost of transmission plus the foregone profit contribution). This is trivial for homogeneous Bertrand competition. In cases of heterogeneous goods and other types of competition it would mean that the incumbent would expand in the wholesale electricity market up to the point where the marginal profit contribution from more wholesale electricity sales equals that from more sales to independent generators. Thus, the ECPR is conjectured to appear as an equilibrium result of competition rather than as a starting point of transmission price setting by an incumbent with market power.[23]

Restricting regulation to transmission, while leaving generation virtually unregulated, is certainly bolder than having (separate) regulation for both stages. However, both approaches appear to be workable.

6. Conclusions

We have suggested a price-cap approach for independent transmission companies. This approach would be implemented for two-part tariffs where the variable part would reflect congestion charges, power losses and other ancillary services while the fixed part would reflect capacity costs. The firm would then have incentives to trade off congestion against capacity expansion in such a way that it becomes profitable to expand, whenever the costs of congestion on average exceed the costs of expansion. With sensible parameter values the scheme should fulfill the six principles identified by Green (1997) for the design of transmission prices.

For integrated generation and transmission companies facing competition in generation we suggest either separate price-caps for generation and transmission or only price-caps for transmission, with generation unregulated (only subject to antitrust laws). In either case, an imputation rule would be added, assuring that the integrated

[23] I first published this conjecture in Vogelsang (1996). Stronger statements have been made by Tye, who maintains that the ECPR will always result *ex post*. To the best of my knowledge no general proof of the proposition has appeared in the literature.

company charges transmission to itself at the same rates available to outsiders.

The price-cap schemes discussed for electricity transmission can be easily adapted to the regulation of natural gas and oil pipeline companies. Complications could arise from the fact that oil and gas are substitutes in demand and complements in supply. Simpler versions of price-caps could hold for gas or electricity distribution companies. In the latter two cases the price adjustment should be based on general inflation, and the X factor should include the difference between expected inflation and those items not covered by an input price adjustment factor.

References

Armstrong, Mark, Chris Doyle and John Vickers (1996), "The Access Pricing Problem: A Synthesis", *Journal of Industrial Economics*, 44, pp. 131-150.

Baumol, William J. (1983), "Some Subtle Issues in Railroad Regulation", *Journal of Transport Economics*, 10, pp. 1-2.

Baumol, William J. and J. Gregory Sidak (1994), *Toward Competition in Local Telephony*, Cambridge, MA: MIT Press and American Enterprise Institute Press.

Bertoletti, P. and C. Poletti (1997), "Welfare Effects of Discriminatory Two-part Tariffs Constrained by Price-Caps", *Economics Letters*, 56, pp. 293-298.

Braeutigam, R. R. (1980), "An Analysis of Fully-Distributed Cost Pricing in Regulated Industries", *Bell Journal of Economics*, 11, pp. 182-196.

Brown L., M. A. Einhorn and I. Vogelsang (1991), "Toward Improved and Practical Incentive Regulation", *Journal of Regulatory Economics*, 3, pp. 313-338.

Cabral, L. M. B. and M. H. Riordan (1989), "Incentives for Cost Reduction Under Price-Cap Regulation", *Journal of Regulatory Economics*, 1, pp. 93-102.

FCC, Federal Communications Committee (1996), "Implementation of the Local Competition Provisions in the Telecommunications Act of 1996", First Report and Order, CC Docket 96-98, FCC 96-325, issued August 8.

Fraser, R. (1995), "The Relationship Between the Costs and Prices of a Multi-Product Monopoly: The Role of Price-Cap Regulation", *Journal of Regulatory Economics*, 8, pp. 23-31.

Gilbert, R. and D. Newbery (1988), "Regulation Games", Discussion Paper No. 267, Centre for Economic Policy Research, September.

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

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

Joskow, P. L. and R. Schmalensee (1983), *Markets for Power*, Cambridge, MA: MIT Press.

Laffont, Jean-Jacques and Jean Tirole (1993), *A Theory of Incentives in Procurement and Regulation*, Cambridge, MA: MIT Press.

——— (1994), "Access Pricing and Competition", *European Economic Review*, 38, pp. 1673-1710.

——— (1996), "Creating Competition Through Interconnection: Theory and Practice", *Journal of Regulatory Economics*, 10, pp. 227-256.

Loeb, M. and W. A. Magat (1979), "A Decentralized Method of Utility Regulation", *Journal of Law and Economics*, 22, pp. 399-404.

Lyon , T. P. (1991), "Regulation with 20-20 Hindsight: 'Heads I Win, Tails You Lose'?", *RAND Journal of Economics*, 22, pp. 581-595.

Masmoudi, Hautam and Francois Prothais (1994), "Access Charges: An Example of Application of the Fully Efficient Rule - Mobile Access to the Fixed Network", mimeo, April .

Neu, W. (1993), "Allocative Inefficiency Properties of Price-Cap Regulation", *Journal of Regulatory Economics*, 5, pp. 159-182.

Pérez-Arriaga, I. J., F. J. Rubio, J. F. Puerta, J. Arceluz and J. Marin (1996), "Marginal Pricing of Transmission Services: An Analysis of Cost Recovery", in M. Einhorn and R. Siddiqi (eds.), *Electricity Transmission Pricing and Technology*, Boston/Dordrecht/London: Kluwer Academic Publishers, pp. 59-76.

Sappington, D. E. M. and D. S. Sibley (1992), "Strategic Nonlinear Pricing Under Price-Cap Regulation", *RAND Journal of Economics*, 23, pp. 1-19.

Schweppe, F., M. Caramanis, R. Tabors and R. Bohn (1988), *Spot Pricing of Electricity*, New York: Kluwer.

Sibley, D. (1989), "Asymmetric Information, Incentives and Price-Cap Regulation", *RAND Journal of Economics*, 20, pp. 392-404.

Vogelsang, I. (1989), "Two-Part Tariffs as Regulatory Constraints", *Journal of Public Economics*, 39, pp. 45-66.

——— (1990), "Optional Two-Part Tariffs Constrained by Price-Caps", *Economics Letters*, 33, pp. 287-292.

Vogelsang, I. and J. Finsinger (1979), "A Regulatory Adjustment Process for Optimal Pricing by Multiproduct Monopoly Firms", *Bell Journal of Economics*, 10, pp. 157-171.

Willig, Robert D. (1979), "The Theory of Network Access Pricing", in Harry Trebing (ed.), *Issues in Public Utility Regulation*, East Lansing: Michigan State University, pp. 109-152.