

# **PBR Designs for Transcos: Towards a Comparative Framework**

**Steven Stoft and Frank Graves\***

June 9, 2000

<b>Introduction</b> .....	1
<b>A Fundamental Regulatory Tradeoff</b> .....	2
<b>What Is Transco PBR?</b> .....	3
<b>Three PBR Schemes</b> .....	6
<b>Using Congestion Prices Alone</b> .....	6
<b>Using Throughput: Price Cap Regulation</b> .....	7
<b>Using Congestion, Loss and Expansion Costs</b> .....	7
<b>PBR Scheme No. 2: Price-Cap Regulation</b> .....	9
<b>Price-Capping Flow on Wires</b> .....	10
<b>Price-Capping Delivered Energy</b> .....	10
<b>Price-Capping a Flow-Energy Bundle</b> .....	11
<b>PBR Scheme No. 3: The "Ideal" PBR</b> .....	13
<b>Problems with the "Ideal" Incentive</b> .....	14
<b>Scaling the Problem Down</b> .....	14
<b>Dynamically Adjusting the Price Cap</b> .....	15
<b>Back to COSR</b> .....	16
<b>Conclusion</b> .....	17

---

\* Steven Stoft is a Senior Advisor to and Frank Graves is President and CEO of *The Brattle Group*. The authors would like to thank Ingo Vogelsang, William Hogan, Paul Joskow, Peter Fox-Penner, and Johannes Pfeifenberger for many helpful corrections, insights and discussions, but are solely responsible for remaining flaws.

## Introduction

The recently promulgated FERC Order 2000 envisions a new form of organization administering electric transmission functions, the RTO, for Regional Transmission Organization. All public utilities under FERC jurisdiction, except those in existing ISOs, must file an RTO plan with FERC by October 15, 2000, and are to participate in an approved RTO by December 15, 2001. Such an RTO is expected to implement congestion management by December 15, 2002 and transmission planning and expansion functions by December 15, 2004. The October 15<sup>th</sup> filing need not propose an RTO, but may instead simply explain the impediments to doing so.

In Order 2000, the FERC uses the term RTO to include both ISOs and transcos, and it discusses performance-based ratemaking (PBR) for RTOs. Order 2000 devotes considerable attention to the importance of PBR to future RTO pricing and performance. The basic notion is that Cost of Service Regulation (COSR, hereafter) may no longer be adequate for motivating the kind of active, regional market facilitation that FERC now expects RTOs to provide. The RTO is directed to manage congestion, loop flows, ancillary services, grid expansion and generation siting through pricing and other policies. Many of these involve costs and circumstances external to the cost of simply owning and maintaining wires, which is all that COSR typically compensates. Thus, it is likely that new pricing flexibility and sophistication will be essential. PBR is regarded by the FERC, and many commentators on Order 2000, as a platform for this needed sophistication. However, PBR theory as currently developed applies only to for-profit organizations, and it has never been applied in the U.S. to unbundled transcos. The FERC is contemplating extensions of this theory to non-profits, such as ISOs, but this article will not. It will examine only for-profit transcos.

Transcos provide quite a number of services to their customers: power delivery, grid reliability, transfer capability expansion, loss management, voltage support, and coordination with other neighboring transcos via Area Control Error management, line loading relief cooperation, and various information services about the state of the regional system. In addition, they may run power exchanges and provide a host of associated accounting and settlement services. Of these many functions, the transport of bulk power from generators to distribution networks, which we will call simply “transmission,” is the most fundamental service. This is also the most difficult service to measure and thus to regulate successfully. If the problem of PBR for this basic transmission service can be solved, the techniques for regulating the other services should be easily discovered.<sup>1</sup> Consequently, this article focuses on the problem of designing a PBR for transmission service alone.

---

<sup>1</sup> Possibly with the exception of coordination services, which involve complex externalities. Ultimately, the obligation to coordinate for large regional grid reliability may be simply a regulatory requirement, with no market or profit incentives to elicit it.

The goals of this article are modest. We do not attempt to solve the problem, but only to illuminate it. This article collects a few of the most commonly mentioned approaches to transco PBR and tests them in simple networks, or subjects them to a modest level of economic scrutiny. The goal is to demonstrate the nature of the transco PBR design problem and some techniques for testing solutions. As a byproduct we present several useful facts about some proposed PBRs.

## **A Fundamental Regulatory Tradeoff**

The possibility of continuing to regulate transmission under COSR is barely mentioned by the FERC in Order 2000, while PBR is discussed extensively. This may be just devoting attention to what is new and flexible under the Order, but the resulting distancing from COSR is probably wise. Transcos would be more difficult to regulate efficiently with COSR than were vertically integrated utilities. There are two sources of this difficulty: first transmission is a notoriously difficult product to measure. Having a means of measuring the quantity and quality of transmission service provided is critical before notions of prudence (or “used and usefulness”) can be applied to approving capital expenditures and costs in order to set COSR rates. Second, the FERC has now clearly required efficient congestion pricing. Congestion pricing is an important step towards efficient regulation, and it is not particularly difficult to implement under COSR. But once implemented, it makes the “used and usefulness” test difficult to apply. This is because congestion pricing requires that flow on an underutilized line be charged zero \$/MWh for its use of the line. This same price would apply to a line that is useful but slack and to one that is considerably overbuilt. Moreover, this price is no reflection of the cost of the line, nor is the congestion price on a constrained line. COSR of generation was much easier. Energy from plants operating below full capacity was never priced at zero, and it was not difficult to spot excess capacity.

To understand the difficulties in choosing a transmission-pricing regime, it is helpful to review a fundamental regulatory trade-off: efficiency versus rent extraction. In economics, the term “rent extraction” means having customers pay as little above the cost of service as possible. The traditional view, still current in many jurisdictions, is that the regulator’s goal is simply to eliminate such rents, at least under expected operations.<sup>2</sup> The modern academic view, *e.g.*, as expressed by Laffont and Tirole, is that 100 percent rent extraction is undesirable because it leads to 0 percent efficiency. In other words, if regulators were to achieve their goal of perfect (and unconditional) COSR, it would result in poor performance because all incentives to improve would have been eliminated. Fortunately regulators have pursued COSR and rent extraction less strictly, in a way that has often been beneficial to the efficiency side of the fundamental tradeoff.

---

<sup>2</sup> The motivation is that such strictness keeps allowed costs, hence rates, down for customers, and that additional profits (beyond the cost of capital) are unnecessary, unmerited, and perhaps even unfair.

As many industry analysts have pointed out, COSR in practice is a special case of price-cap regulation: ratecases set the price of power for a considerable period of time, typically about three years. The intention is to set price so that initially the utility recovers no more than its projected cost of service, which is estimated by adjustments to recent past actual costs. Once the price is set, the utility is typically allowed to keep any extra profits that it makes by reducing cost. Thus, between rate cases, COSR is like RPI – X price-cap regulation, albeit with X equal to RPI. The incentive provided by this intermittent form of price-cap regulation is less than full strength, because the firm anticipates that its long-run cost savings will be re-captured by the regulator at the next rate case. This demonstrates the fundamental regulatory trade-off: More frequent rate cases allow closer tracking of the cost of service, but they dampen the regulated firm’s incentive to cut costs.

The second of the three incentive mechanisms discussed in this article is price-cap regulation. Because realistic COSR is implicitly a form of price-cap regulation, we will not analyze COSR separately, but instead argue that it suffers from the same defects as explicit per MWh price-cap regulation. For the moment we will only note that congestion pricing prevents the regulator from setting the commodity price of the service that is delivered by the transco. This is because congestion pricing reflects differences in short run marginal costs of power between adjacent nodes on a grid. Congestion fees so designed typically recover only around 20 percent of the embedded cost of transmission service, but they cannot be scaled up for the missing 80 percent because doing so would distort efficient price signals and violate the FERC’s requirement of efficient congestion pricing. This means some other source of revenue must be found, such as an energy charge imposed on power injections without regard to flow congestion. But energy is not the transco’s product, and capping a charge on someone else’s product (the generator’s) is not the same as capping the price of the regulated firm’s product. (In practice, pools with nodal pricing have handled this by rebating the transmission congestion rents (TCRs) to the transco owners and recovering the transco revenue requirements through access charges assessed against the TCR holders. Thus congestion pricing can be implemented under COSR, but not in conjunction with one-part pricing as often obtains under price cap regulation.)

### **What Is Transco PBR?**

The FERC devotes several pages to transco PBR but gives only broad hints of how it should work. For instance, various structural features of PBR are recommended, such as devising terms that yield risks and rewards to owners while sharing the benefits of improvements with customers. Different broad mechanisms are acknowledged as potentially acceptable, such as UK-style price-caps or yardstick competition between similar RTOs. However, there is no guidance on design of the indices or specification of the performance targets. For instance, there is no suggested measure of grid reliability to be served by the transco, probably because there is no simple standard. Of course, specifying a strict PBR formula was not the purpose of Order 2000, nor does the FERC desire to

foreclose creative proposals of any kind. However, it turns out that specifying an effective formula is difficult, which is the main point of this article.

As noted above, this article will focus strictly on transmission service proper, as this is both the fundamental service of a transco and the most difficult to regulate. This service is provided primarily by wires and transformers, and for simplicity we will focus on the use and expansion of wires for the purpose of trading power economically. The wires are not useful or interesting for their own sake, so incentive regulation must focus on how they can be used to improve performance of adjacent markets for which the wires are essential facilities. This highlights another significant drawback of COSR as well as conventional price-cap PBR: they both focus solely on the direct costs of owning and maintaining wires, and not at all on the indirect costs and benefits of how those wire management practices affect upstream and downstream markets. A proper transco PBR scheme must reflect these externalities, which can be quite large relative to the direct costs of the wires themselves.

As a benchmark for testing PBR schemes we need to know what an ideal transco can and should do to improve the efficiency of power trading and delivery. Much of the opportunity to improve these areas will arise from the provision of additional transmission lines. Increased transfer capability could play a critical role in improving the reliability and competitiveness of wholesale power markets, now generally recognized as inherently prone to occasional exercises of market power by generators.<sup>3</sup> Transcos will soon be at the forefront of this growing problem, and PBR incentives may be a key to its solution.

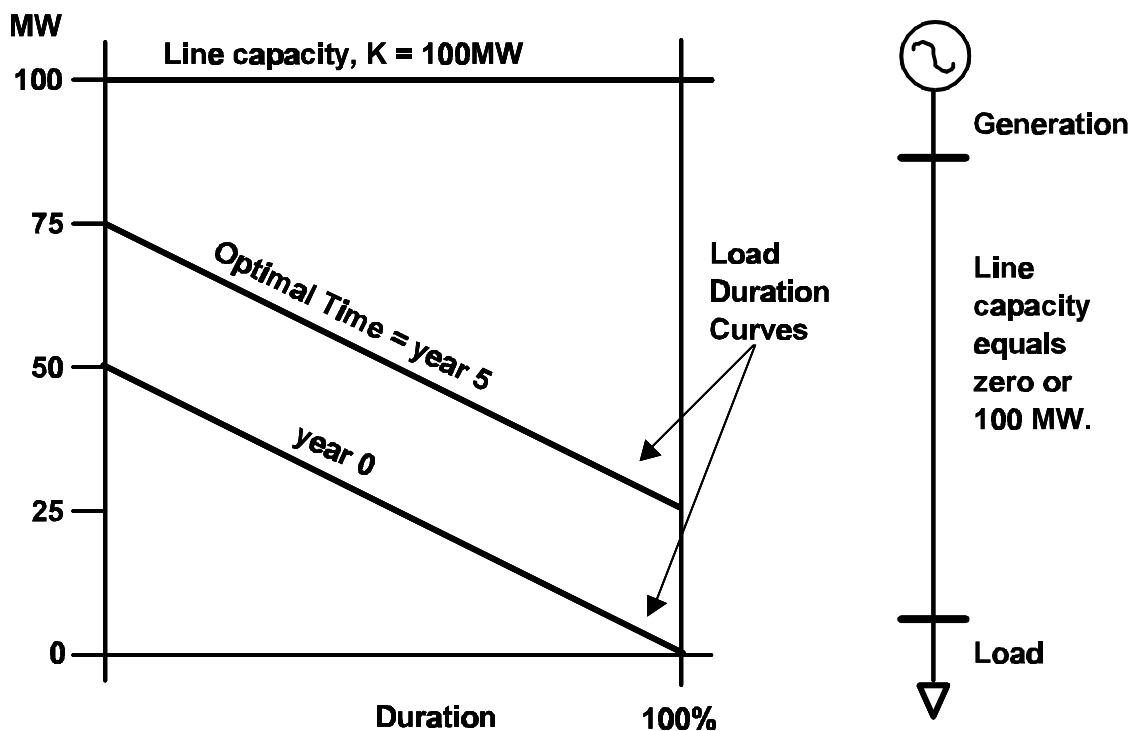
Thus to test any PBR scheme we need a model of an efficient line upgrade. In general this problem is extremely difficult because it involves computing savings of generation costs over long time horizons. Note that these costs are external to the transco operations and accounts, immediately flagging a difficulty for designing a desirable PBR system! We need to make the transco responsible, in part, for the success and costs of an independent sector. Fortunately there is one restricted investment problem that is relatively simple to pose without needing detailed generation cost forecasts, and it will demonstrate the PBR dilemma well enough.

When a line is built, two decisions must be optimized: timing and line size. The harder one is line size, because it involves future cost savings. Once optimal line size has been determined, then the

---

<sup>3</sup> According to *The Wall Street Journal*, (April 15, 2000, p. A6), Secretary of Energy Bill Richardson was told by state officials at three regional "summit" meetings that the government needs to provide more incentives for transmission lines, otherwise the new generation coming on line will do little to improve reliability. Richardson expressed concerns about the existence of generation market power in a recent speech to NARUC, documented in a March, 2000 report from the DoE Office of Policy, entitled "Horizontal Market Power in Restructured Electricity Markets."

line should be built as soon as its current savings are as great as its amortized cost.<sup>4</sup> Consequently a very simple investment problem results if we assume we know which line should be upgraded, the future cost savings from an upgrade, and that the choice of line size can be restricted to a single capacity. Given all this, the only remaining choice is when to build. A good PBR system should at least be able to motivate good expansion timing decisions.



**Figure 1: Optimal Line Investment**

To model this choice, consider a system with load at one node and cheap generation at another. Assume a certain fixed savings of \$10/MWh of power purchased from the cheap generation. Power can always be purchased at the load bus, but at a \$10/MWh higher cost. The load duration curve at time zero (year = 0) and at the end of year five (year = 5) is shown in Figure 1. The cost of the line, of capacity  $K = 100$  MW, is \$5/MWh. (Typically the cost of line capacity is given in \$/year/MW, but this can be converted to the more convenient \$/MWh by dividing by 8760.) Under these assumptions, the line should be built as soon as it would be half used, which happens at the end of year five as shown in Figure 1. Maximum demand is assumed to increase continuously at 5MW/year. The growth has a 100 percent load factor, solely for the convenience in the calculations. By year 5, the new line would have a 50 percent load factor, saving \$10/MWh, which covers its own

<sup>4</sup> This view is a little simplistic because the two problems usually must be solved together, but this explanation is not misleading in the present context.

\$5/MWh annualized cost at a 100 percent load factor.<sup>5</sup>

We will use this example to see if a transco would actually choose to expand on the same, optimal schedule under some PBR alternatives.

### **Three PBR Schemes**

We will examine PBR schemes that are based on three-performance measure: congestion prices, throughput, and congestion/loss costs. In brief, those based on congestion prices alone fail because this measure is too parochial and creates perverse incentives.<sup>6</sup> Those based on throughput fail because there is no appropriate definition of throughput. Those based on full transco exposure to all congestion, loss, and expansion costs show the most potential in that they can produce ideal transmission incentives if congestion/loss costs are known and generator location is exogenous. But both of these conditions are problematic. Moreover, this method is prone towards overpayment of transcos. A diluted form of this approach may be the pragmatic solution.

#### **Using Congestion Prices Alone**

The idea that congestion prices can somehow be used for transco incentives is common. Sometimes it is suggested that transcos be rewarded for congestion charges and sometimes that they be punished for congestion charges. While rewarding the transco with congestion rents has the obvious problem of encouraging them to make a bad situation worse, this proposal does have the desirable property of getting some lines built eventually. This scheme simply allows the transco to act as a monopolist, subject to the restriction that once a line is built its capacity cannot be withheld (except by pretending there is a physical problem).

The idea of penalizing the transco for the amount of the congestion rent is sometimes mentioned, but does not seem to have been seriously proposed. This may be because there are two ways for a transco to reduce congestion costs to zero: build a high-capacity line or restrict scheduling across the interfaces that trigger congested lines, effectively taking them (partially) out of service. Taking the lines out of service tends to be cheaper, though this might be prevented by reliability incentives, or fixed cost-recovery disallowances. The key point is that optimal congestion is not zero congestion. In many circumstances, the re-dispatch costs of adjusting to an occasionally constrained

---

<sup>5</sup> Note that the present value (PV) of future savings would “justify” building the line earlier, but there would be no advantage to doing so, as in the earlier year the line would be a money loser. That is, the PV of savings is largest if the line is built for year 5. Working with the annualized cost allows an easy test of the optimal timing.

<sup>6</sup> We assume congestion prices are differences between nodal marginal costs as calculated by a DC optimal power flow, as in PJM, NYPOOL, and other LMP-based pools. Such optimization ignores losses.

interface are lower than line (or generation) expansion costs. Indeed, this is one reason why there are “load pockets” enclosing many major cities worldwide. A transco is not serving societal welfare by either eliminating all such bottlenecks or denying access to them.

### **Using Throughput: Price Cap Regulation**

A second proposal, frequently mentioned at the FERC’s technical conference on ISOs, is the idea of using “throughput” as the measure of performance and of rewarding it accordingly. (Congestion is neither rewarded nor penalized, except implicitly.) This seems to have more merit, in that a “good” rather than “bad” is being rewarded, but the meaning of the term “throughput” proves illusory. We will examine two candidates for the definition of throughput: the amount of power delivered and the total MWh-miles clocked in the process of delivering that power. The problem here again will be that throughput can be increased with grid adjustments that do not have favorable societal cost/benefit ratios.

### **Using Congestion, Loss and Expansion Costs**

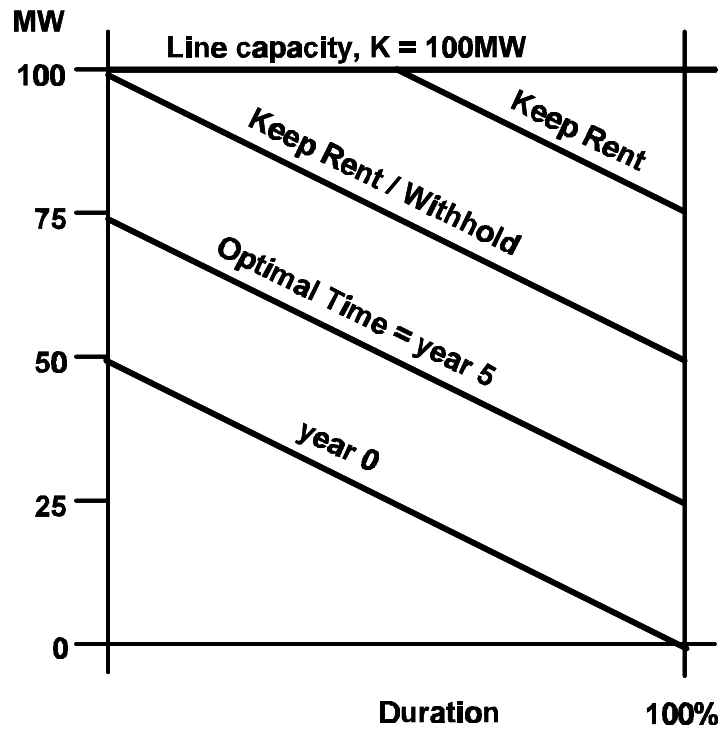
A third possible PBR is proposed by standard economic theory: have the transco internalize all aspects, including both costs and benefits, of when and how to pursue expansion to reduce congestion and losses. This has the advantage of providing the ideal incentive, but the disadvantage of perhaps having to allow too much transco profit in order to avoid the even worse mistake of putting the transco out of business. However there are a couple of directions in which to seek improvements that may eventually yield a solution. The basic idea is to make the transco responsible for all costs caused by any imperfection in the transmission network. If the re-dispatch costs due to congestion and losses are \$X/h, then the transco has those costs deducted from its total revenue. As with any price-cap method it is also responsible for the costs of its investments in the grid, but it could not “rate base” those costs. As a consequence, the transco will seek to minimize the sum of grid imperfection costs plus the cost of building the grid. It will retain the full benefit of improvements in those costs. This causes it to make exactly the right tradeoff between the two costs. (This conclusion takes generator location as a given, and it assumes that grid will only be optimized relative to the given generation set. The problem of inducing correct generator location is even more difficult than the problem of transco incentives.)

Each of these three PBR alternatives is critiqued below, using our expansion timing problem or a related example as a test of usefulness.

### **PBR Scheme No. 1: The Transco Keeps the Congestion Rent**

Consider a PBR scheme that allows a transco to keep all of the congestion rent, and this is its only source of revenues. Such a transco will not aspire to maximize congestion, any more than a

monopolist tries to maximize price. Instead, it will balance throughput, expansion, and congestion to maximize profits. In keeping with the FERC’s direction on this matter, we will assume that congestion is priced efficiently. Typically, when a line is built at the optimal time and of the optimal size, it is initially free of congestion, or is congested only in very unusual circumstances. This is because once fixed infrastructure is in place, it is relatively cheap to increase line capacity. If it remains congested once built, it probably should have been built sooner. For a transco under congestion-rents PBR, it is not profitable to build a line that will initially collect no rent, so either too small a line will be built, or it will be built much too late.



**Figure 2: Optimal Line Investment**

In our simple example, with a fixed-size expansion line it turns out that the line would be built ten years later than the optimal time. Line costs and savings are the same as before: congestion rent is \$10/MWh and the cost of the line is \$5/MWh. With these values it is socially beneficial to build the line when it is half used. But a transco is focused on profit, not social benefit, so a transco that is rewarded with congestion rent will build only when congestion rent equals the amortized line cost. Since the line is congested only when the load-duration curve is at or above the line capacity, Figure 2 shows that the line will not be congested half of the time until the load-duration curve reaches the level shown by the line labeled “Keep Rent.” At this point in time the transco can collect a rent of \$10/MWh half the time and \$0/MWh half the time, for an average of \$5/MWh, which is exactly equal to the line’s cost. From this point on, as congestion increases, the line will actually become profitable. Because the load-duration curve is moving up 25MW every five years, our rent-seeking transco builds

the line at year 15, ten years too late.

The disincentive for the transco under this scheme is that once it builds the line it must make all of its capacity available to the energy market. This is of course the efficient way to run the grid. But because of this rule the transco postpones building the efficient line. If the owner could withhold some capacity, it would build sooner, by five years in our example. To see this, consider the situation at the end of year 10. The load duration curve has shifted up to the point where line use never falls below 50 percent and sometime rises to 100 percent. This is shown by the line “Keep Rent/Withhold.” At this level the line generates no congestion rent if use is unrestricted, but it would generate a revenue of \$500/h if it had a capacity of only 50MW. This is the maximum revenue that can be generated by withholding and is just sufficient to pay for the line. If the transco is allowed to withhold line capacity it will build the line at the end of year ten and will withhold half of its capacity. As load grows it will withhold less and less, always making sure that the declared capacity is fully utilized 100 percent of the time. In this way it will generate maximum revenue.

Although the Keep Rent/Withhold form of regulation gets the line built much sooner, the withholding of capacity reduces the social benefit of the line. Consequently it is impossible to say in general if this approach is better or worse from a social perspective. This is an area for future research as such schemes are under active consideration and little is known of their economic properties.

## **PBR Scheme No. 2: Price-Cap Regulation**

The concept of “throughput” was mentioned by six speakers at the FERC’s ISO conference in the spring of 1998. The use of it in transco incentives was explained most explicitly by Mr. Vesey, Vice President of the Transmission Business for Entergy Services, Inc., one of the first to file with the FERC for a permission to set up a transco.<sup>7</sup> He said:

While this company would provide non-discriminatory access, it would also be driven through appropriate incentives to minimize costs, maximize throughput, achieve efficient levels of congestion and reliability, and expand the transmission grid when economically justified.

Mr. Roy Thilly, CEO of Wisconsin Public Power, Inc. offered similar views, “I think that ISOs can be incentivized, as can transcos, by compensation based on throughput, utilization of the system.” These ideas are widespread and should not be attributed to these individuals in particular.

The idea of “throughput” incentives is common. Unfortunately, as will be described shortly,

---

<sup>7</sup> Inquiry Concerning The Commission's Policy on Independent System Operators Before the Commissioners, Federal Energy Regulatory Commission, 888 First Street, N.E. Room 2 C, Washington, D. C. Wednesday, April 15, 1998.

two meanings of throughput have been confounded. The first meaning is aggregate flow on wires, not differentiated by direction. This is perhaps the natural meaning to use when price-capping a transco. The second meaning is delivered energy, which presupposes a specification of end-use customers attributed to the territory of each transco. Delivered energy is a joint product of generation and delivery, and its principle component, energy, is not produced by the transco. Some price-cap proposals place a price cap on a bundle of goods consisting of both flows on wires and delivered energy. All three will be considered.

### **Price-Capping Flow on Wires**

The most straightforward approach would seem to be to put a price cap on the transco's product, which is essentially the transfer of power. This could be measured by MW miles of power flow if a summary measure were desired, or it could be measured on each wire individually and a different price and price cap could be associated with each wire. The problem with any such method, when used as the sole method of revenue collection, is that it necessarily contradicts congestion pricing, and, as mentioned earlier, the contradiction is dramatic. Congestion prices collect far less than the total revenue requirements associated with the grid. Any flow-based method that collects enough revenue will inevitably charge far more than optimal congestion prices. Consequently any per kWh price-cap method that is strictly flow-based will inevitably cause significant inefficiencies in the use of the grid.

### **Price-Capping Delivered Energy**

An obvious source of funds to cover revenue requirements is a simple charge on delivered energy. This is in fact effectively what is being done in the California, PJM and New York to cover the fixed cost of wires as well as the costs of running their ISOs and procuring ancillary services. To convert this approach to a price-cap incentive, one would simply fix the charge on delivered energy at, say, \$5/MWh and then let the transco build lines as it saw fit. Obviously if the initial price-cap just allows the transco to recover its costs, it will do well by doing nothing. Say demand grows three percent per year. The transco's revenues will also grow three percent per year, and all of this will be profit, assuming there are no congestion costs that the transco must absorb. But price caps typically use an RPI - X formula, so it would be possible to adjust X to approximately remove this growth-induced windfall gain. What incentive would this mechanism provide? If the transco upgrades a transmission line, when will that increase its revenues? It will do so if and only if building the line increases the consumption of power. But building wires generally has little effect on power consumption. Loads generally have full and reliable access both before and after any particular upgrade, so they will increase their power usage only if the price goes down. Thus the transco would be rewarded only to the extent that the new line reduced the price of power and thereby caused consumers to buy more. If demand were very sensitive to price, this effect might be noticeable, but this is unlikely to be the case, so this approach would result in an inadequate rate of transmission

expansion.

### Price-Capping a Flow-Energy Bundle<sup>8</sup>

In its richest form, modern price-cap theory and practice involves capping bundles of goods, *i.e.*, capping a price index based on a group of goods produced by the regulated firm. This allows the firm some flexibility in setting relative prices while keeping the average price level in check. This can be illustrated on our two-product bundle of transco goods as follows. Let  $F$  stand for the quantity of flow purchased by consumers, and  $E$  stand for the amount of energy purchased. We do not need to worry yet about exactly how flow is measured; it is only necessary to assume that this is possible and that it can be priced on some per-unit basis. Price-capping a bundle works by determining a revenue  $R_0$  at some time 0 and measuring the actual consumption of goods at that time. Say that  $F_0$  and  $E_0$  are the quantities consumed at time 0. The price cap for period 1 says that the firm (transco) must not set prices in such a way that  $F_0$  and  $E_0$  together cost more than  $R_0$ . That is, they must not cost more than they did originally. Algebraically this is stated as:

$$P_1 H F_0 + p_1 H E_0 \# R_0,$$

where  $P_1$  is the price of flow and  $p_1$  is the price of energy, each in period one.<sup>9</sup>

This price-cap rule allows the transco to pick from a whole family of different price pairs. It can pick a high-energy price and low flow price or just the reverse. Whatever it picks, consumers will make their choice and this will determine some  $F_1$  and  $E_1$  quantities in period 1. These quantities determine a new revenue,  $R_1$ , for the transco. These new quantities and the new revenue are then used with the same price-cap rule to limit the transco's pricing in period 2. One benefit of this method is that consumers (in aggregate) can never be made worse off by any change in price. Another benefit is that the transco can move prices in line with marginal production costs.<sup>10</sup>

While this price-cap methodology seems more aimed at finding the right prices than at getting the transco to build the right lines, it has been proposed for doing both at once. Because congestion prices are "the right flow prices," this would mean that if such a method worked, it would solve the transmission investment and re-dispatch problems simultaneously. That is the central goal of transco

---

<sup>8</sup> For a more complete and rigorous discussion of this topic see Ingo Vogelsang, "Price Regulation for Independent Transmission Companies" (Boston University mimeo, September 1999).

<sup>9</sup> This definition of the price cap is based on a Laspeyres price index. It is also possible to base on a price cap on a Paasche price index. The results will differ.

<sup>10</sup> If completely unconstrained, this price-cap method leads the transco to Ramsey prices, *i.e.*, inverse elasticity differentiation of the prices for the two goods, which are the most efficient prices given a capped amount of revenue that can be collected. In practice, bundled-good price caps may place restrictions on the rate or maximum degree of change in the component prices.

regulation.

For the price-cap method to succeed at this,  $P$ , the price of power flows, must be the set of congestion prices. This means that the price cap does not just control two prices,  $P$  and  $p$ , but that the  $P$  is actually a very long list of prices. Theoretically this is not a problem. Congestion prices are being calculated in PJM and elsewhere, and the same price cap formula that works for a basket of two prices readily expands to handle any number of prices. As noted above, one idea is to have an RTO transco compute the correct congestion price vector  $P$  and to pay the transco the whole congestion rent  $P \cdot F$ .<sup>10</sup> This congestion rent is quite insufficient to cover the transco's revenue requirement, so the charge,  $p$ , on energy could be used to supplement the congestion charge. The price cap formula would be applied as above, but with the  $P_1 H F_0$  replaced by the computation of a pseudo congestion charge base on last period's flows and this period's congestion prices.

How would such a scheme work in operation? The transco's revenue requirement would be chosen as the initial  $R_0$ . The initial congestion prices would form the price vector  $P_0$ . The initial energy charge  $p_0$  would be found by subtracting the initial congestion charge and dividing by  $E_0$ , the initial level of delivered energy. During the first year of operation, the RTO transco would compute new congestion prices,  $P_1$ , and use them in real time. It would also record the actual flows  $F_1$ , and the delivered energy,  $E_1$ , for future use in setting the energy price  $p_2$ . At the end of the year a new energy price,  $p_1$ , would be computed according to the price-cap rule as described above:  $p_1 = (R_0 - P_1 \cdot F_0) / E_0$ .  $P_2$  would again be set at actual congestion rates. This process would continue indefinitely, with occasional adjustment if the transco's revenues were getting too far out of line in either direction. (Note that this approach assumes the transco does not actually bear the congestion costs. It passes them on to customers via nodal prices and does *not* rebate the congestion rents to FTR holders. In essence, the transco is *the* sole FTR holder. Of course, the FTRs are of no particular use to the transco, so it would probably try to sell them to energy traders at a profit.)

What investment incentives would be provided by such a scheme? If the transco upgraded a line, would it be rewarded? Consider a single congested 100 MW line with a constant congestion price of \$20/MWh. Say the transco expands the line by 100 MW and thereby lowers the congestion price to zero. Because of the price cap formula, the decrease in congestion rents would be entirely made up by an increase in subsequent energy charges, but this only leaves the transco breaking even. It gets the same amount of revenue as it would have without building the line; it is not rewarded at all with a piece of the net social benefits it creates. On the other hand, if it expands the line by only 50 MW, and that lowers the congestion rent only to \$10/MWh, the loss of congestion rent on the pre-

---

<sup>10</sup> Using the dot for vector multiplication denotes multiplying each price by the corresponding power flow and then summing the results. This gives the total congestion charge, also called the congestion rent.

existing part of the line will be entirely made up by the increase in the energy charge. But this time there will be additional revenue of \$10/MWh on the 50 MW of new line. This will reward the transco for building the line, but it will not do so fully enough. It can be calculated that this reward, \$500/MWh, is less than the social value of the line, which, assuming a linear demand curve, is \$750/MWh. But this is not so far off, suggesting that this approach has some promise.<sup>11</sup>

Several conclusions can be drawn regarding price-cap regulation and throughput incentives for transcos. First, a price cap with any hope of success will be quite a complex mechanism involving a large number of prices and the calculation of efficient congestion prices. Second, no off-the-shelf mechanism is currently available. Instead, this is an area in need of economic research. One final point is more optimistic. The above mechanism provides an approximation of the decrease in re-dispatch costs attributable to the upgrade. If this cost decrease were just calculated more accurately and given to the transco, that would provide the ideal incentive. That is exactly what the next mechanism does. It simply tackles this calculation and reward problem directly.

### **PBR Scheme No. 3: The “Ideal” PBR**

The concept behind the “ideal” PBR is to put the transco at risk for the grid “imperfection” costs arising from congestion, losses, outages, voltage drops, and the like. The scheme does provide nearly ideal incentives, in the sense of encouraging efficient and timely improvements to the grid, but it makes it difficult to control the transco’s profit level.

The first step in this scheme is to compute the cost of congestion. Notice that during the first five years of our simple expansion-timing model above, there is an implicit cost of congestion. Because there is no line, the expensive local generation is dispatched in place of the cheap remote generation. This is costly, but before building the line no congestion rent is calculated or collected because there is no power flow—again because there is no line. Under scheme No. 1 we were concerned with congestion *rent*. Under scheme No. 3, we are now concerned with congestion *costs*. Obviously the two can be very different. (Congestion costs can also be much less than congestion rents, since the former reflect total costs with and without congestion, while the latter reflect the product of marginal costs and volumes moving between nodes with unequal spot prices.)

Quantifying the total cost of congestion is extremely difficult without a large-regional centralized market, such as the ISOs in PJM, New York, and California. Even with these agencies, quantification can be difficult. The PX in California performs an unconstrained dispatch on the bids it

---

<sup>11</sup> If the price cap had been based on a Paasche instead of a Laspeyres price index, the result would have been over— instead of under—compensation.

receives. This means it computes the dispatch as if the transmission system were perfect, with no congestion and no losses. Of course the PX covers only part of the California market, but ignoring that problem, the cost of energy computed in the PX's idealized world could be subtracted from the actual cost to find the "grid imperfection cost" (GIC). The California ISO would be able to perform this calculation if its incremental and decremental bids were extensive enough.<sup>12</sup> PJM and the NY ISO are in a better position to measure congestion costs, though they currently do not do so. They could compute an unconstrained dispatch solely for the purpose of measuring congestion costs. Although it might appear that it is just a matter of subtracting unconstrained cost from constrained costs, the required procedure is a bit more complex.<sup>13</sup> But given a complete and accurate set of bids, it is quite doable.

Assume now that it is possible to measure the total cost of congestion and losses caused by an imperfect grid, and call that cost GIC, for grid imperfection cost. An optimal incentive scheme is then to allow the transco to collect revenues of some arbitrary but high amount, say \$30 per MWh of delivered power, minus GIC. This number has to be set high enough to ensure that the transco will not go out of business if the GIC should become high. The transco then gets to keep every dollar by which it reduces GIC, so if it can reduce GIC by \$1.00 by spending \$0.95 to improve the grid, it will do so. This is exactly what should be done. Similarly if the cost were greater than the reduction in GIC, it would refrain from making the investment.

### **Problems with the "Ideal" Incentive**

The obvious problem with this scheme lies in setting the \$30/MWh price cap. Certainly this is more than enough for assuring viability, and the result would be an optimally built grid. But this leaves a lot of money on the table. The transco's profits would usually be unconscionably high, and consumers would be much worse off than they were under the old style of regulation. The problem is to set the access price cap as low as possible while allowing the transco at least a normal rate of return. If the cap can be set low enough, the incentive will be a success.

### **Scaling the Problem Down**

The access price cap is set high in the above formulation of the ideal incentive because it must cover the entire cost of the grid plus the variance in CIG. Call  $C_g$  the revenue requirement of the grid

---

<sup>12</sup> The California system has been designed to encourage private congestion management, and so incremental and decremental bidding is less than comprehensive. It would also be necessary for the ISO to include the cost of must-run contracts.

<sup>13</sup> There would be some inaccuracies with this technique due to inaccurate bidding. For example, bids might be deliberately inaccurate to the extent the bidder knew that its bid would not be accepted in the final constrained dispatch. Without a line in place a generator could bid an impossibly large supply in order to make congestion cost look large, or it could not bother to bid at all knowing that its bid would not be accepted in the final dispatch. This problem would require some regulatory scrutiny, but could probably be overcome.

(which does not include any generation costs for congestion alleviation or ancillary services like reserves). It is completely borne by the transco. If we call the access charge  $P$ , and the quantity of delivered energy  $Q$ , the transco's profit is  $PCQ - (GIC+C_g)$ . Say we know that the  $(GIC+C_g)$  cost term will be \$1B/year plus or minus \$400M/year, and we do not want to take any chance of causing the transco to show a negative profit (unrealistic but it illustrates the following point most simply). In this case the price cap must be set so that  $PCQ = \$1,400M/year$ . This leaves the transco with an expected profit of \$400M/year (over and above its normal rate of return on capital), and a guarantee of positive profits. (Excess profits will be between \$0 and \$800M/year.)

There is a cheaper way to achieve this same guarantee while leaving the incentive properties of the ideal scheme intact: Divide the transco's profit formula by two. Because the original profit was known to be positive, half of that profit must also be positive, so the transco is still assured of a positive profit. Because the expected excess profit was \$400M, dividing by two will decrease this value to \$200M, a considerable savings to the consumer. Originally the transco's goal was to minimize  $(GIC+C_g)$ . Its goal now becomes minimization of  $(GIC+C_g)/2$ , but clearly whatever minimizes one also minimizes the other. Therefore, the transco's incentive remains unchanged.

Such a division by two requires some work to implement. Dividing the access charge,  $P$ , in half is easy enough, and charging only half of  $GIC$  is easy. Previously, the grid cost,  $C_g$ , was just left to the transco, and the regulator did not need to keep any account of it. But now, in order to make the transco bear only half of this cost, it is necessary to keep a full accounting. The regulator is now responsible for half of this charge and this half will need to be collected through an additional access charge on top of  $P/2$ . But this charge, with the help of a small balancing account, can be computed easily. The new incentive scheme is to pay the transco  $(GIC+C_g)/2$  and to set the transco's price cap on access at  $P/2$ . Another access charge must be collected to cover the other half of  $C_g$  on the transmission system. These payments and price caps give the transco a profit of  $P/2 + (GIC+C_g)/2$ .

This technique reduces, but does not eliminate the problem. While it might seem that the process could be pushed far enough to make the problem entirely negligible by dividing by say 1000 instead of by 2, there are subtle considerations that might prevent this. These involve the difficulties of the regulator measuring  $C_g$  with complete accuracy. Division by two is certainly sensible, but division by ten or more would require analysis to demonstrate its effectiveness.

### **Dynamically Adjusting the Price Cap**

A second method of reducing the expected excess profit, often referred to as the transco's rent, is to observe the magnitude of this excess and reduce it over time. But this process distorts the ideal incentive. For instance if  $P$  is adjusted so continuously and so effectively that the transco is allowed an excess profit of exactly \$0.10/MWh at all times, then the incentive will have been

completely destroyed. This is because the transco will know that no matter how well or how badly it performs it will make the same excess profit.

The general result is that there remains an unavailable tradeoff between the transco's rent and the power of the incentive. In order to produce a high-powered (ideal) incentive and be sure of keeping the transco in business, the regulator is forced to leave a lot of money (rent) on the table. By reducing the effectiveness (power) of the incentive, it is possible to re-capture some of the transco's rents and to move the transco closer to the risk-appropriate level of profit. Regulation cannot approach both the perfect incentive and zero rents as can perfect competition.

### **Back to COSR**

An alternative approach that has some of the spirit of this ideal incentive but leans much more towards traditional rent extraction down to COSR is as follows. Have the transco collect its revenue requirement on a per-MW-of-access basis. This revenue requirement can grow, as in the past, with rate base additions. Unlike in the past, these would be approved based on present value avoided GICs in excess of the associated expansion costs. That is, the regulators would have to review the full set of future wholesale power costs a transco affects, not just its own direct costs.

What we have here is a solution to the "used and useful" problem that plagues COSR for transcos, but it is not an easy solution. To determine if a line should have been built (will be useful), it is necessary to compute its effect on the present value of the grid imperfection cost (GIC). This is far more difficult than what is required under the "ideal PBR." In that case it was only necessary to compute GIC from real after-the-fact costs. In order to implement this ideal form of COSR, it is necessary to compute GIC far into the future.

Congestion rates would *not* be collected by the transco but rebated to FTR holders, as presently in PJM and elsewhere. However, the transco's allowed return on equity (RoE) would be indexed to reduction in the GIC per MWh of flows. This gives the transco no reason to delay building a socially attractive line simply to let congestion rents fund the investment. To the contrary, it would be penalized (slightly) for delaying, by growth in GIC adverse to its allowed RoE. Likewise, it would not be relying on growth in throughput to justify the expansion outlays, as its revenue requirement assures full cost recovery.

The main drawbacks to this approach are two. First, it requires a great deal of market price and traffic outlook information collection that the Commission(s) could make a compensated responsibility of the transcos. This simply reflects the reality of transco value-added being what must be incentivized, not transco cost reductions. There is no way around this, but this need has not been appreciated in any of the approaches commonly discussed to date for transco pricing, whether COSR

or PBR based on congestion or throughput. The second drawback is that it retains the prospect of Averch-Johnson types of imbalances (positive or negative) in the degree of transco expansion, if allowed rates of return on capital are set too high or low. Given that the transco will be playing a more complicated role than it has in the past (now that it is a market-facilitating intermediary rather than a relatively passive collection of assets associated with generation expansion), agreeing on the fair return may be more difficult. By virtue of participating, albeit in a dampened way, in the risks of the wholesale generation market—generally acknowledged as more risky than any other commodity—an increase in allowed return on equity for transcos may become essential.

## **Conclusion**

We have discovered no performance-based regulation for transcos that appears readily workable, either politically or administratively, let alone ideal. Allowing transcos to keep congestion rents is nothing more than a decision not to regulate, for this is exactly what a monopoly transco would do. Rewarding transcos for throughput has the advantage of rewarding a service instead of rewarding a nuisance (congestion), but if this is to be done well it requires a detailed form of regulation. There is no single \$/MWh-mile value that well approximates the average value of flow on a correctly-sized new transmission line. Because of this, and to do a good job, the regulator would have to evaluate every line individually and often.

Perhaps the most promising approach is some variation of the ideal incentive which makes the transco fully responsible for upgrade costs and for all costs attributable to the grid being less than desired. The problem with this approach is that the transco may have to be given excess profits in order to be sure that it is not put out of business, since congestion costs can be quite large under extreme circumstances, and they are very hard to predict accurately in advance. This is the bane of all price-cap methods. Nonetheless, by making the transco responsible for only a fraction of these costs, a suitable compromise may be discovered.

Finishing the design of the “ideal” PBR or discovering another, better PBR does not appear to be an easy task. This difficulty is present even when only coping with throughput management of congestion and expansion, *i.e.*, ignoring other complex obligations such as procuring ancillaries and monitoring/mitigating market power abuses. If transcos are to be successfully regulated, this problem deserves far more attention than it has received to date.