RESTRUCTURING THE ELECTRICITY MARKET: INSTITUTIONS FOR NETWORK SYSTEMS

WILLIAM W. HOGAN

April 1999

Center for Business and Government
John F. Kennedy School of Government
Harvard University
Cambridge, Massachusetts 02138

CONTENTS

RESTRUCTURING THE ELECTRICITY MARKET: INSTITUTIONS FOR NETWORK SYSTEMS

William W. Hogan[1]

Public policy development for electricity restructuring emphasizes institutions for market operations in network systems. The different models present alternatives for the mix of responsibilities of the necessary system operator. Customer flexibility and choice require efficient pricing; inefficient pricing necessarily limits market flexibility. The analysis points to an integrated independent system operator, de jure on its own or de facto within a larger transmission organization, with locational marginal cost pricing rules, as the model most likely to be successful in preserving system reliability while supporting competitive markets with customer choice.

INTRODUCTION

The process of restructuring wholesale electricity markets in the United States has added to the extensive worldwide debate about the range of possible and preferred alternatives for organizing regional electricity markets. Given the highly interconnected network, it is clear that some aggregation to regional transmission organizations would be necessary. The issues raised are important, but the discussion has tended to focus on ownership and governance questions. These can distract from the more difficult but, in the long run, more fundamental consideration of the rules for market operations within and across regions in network systems. A critical summary of the institutional debate and issues provides the context for a review of the market rules that can support a competitive electricity market.

Policy for the continuing evolution of electricity restructuring should emphasize the institutions for market operations. Interconnections through the transmission grid create the necessity for regional organizations that can accommodate competition in services, generation, and contracting while preserving the reliability of the transmission system. Alternative models are many, but can be grouped under the general headings of "Transcos," "Gridcos," "ISO/PX,"

---

[1] Lucius N. Littauer Professor of Public Policy and Administration, John F. Kennedy School of Government, Harvard University and Principal of the Law and Economics Consulting Group. This paper draws on work for the Harvard Electricity Policy Group and the Harvard-Japan Project on Energy and the Environment. The author is or has been a consultant on electric market reform and transmission issues for American National Power, British National Grid Company, GPU Inc. (and the Supporting Companies of PJM), GPU PowerNet Pty Ltd, Duquesne Light Company, Electricity Corporation of New Zealand, National Independent Energy Producers, New York Power Pool, New York Utilities Collaborative, New England Power Company, Niagara Mohawk Corporation, PJM Office of Interconnection, Putnam, Hayes & Bartlett, Inc., San Diego Gas & Electric Corporation, Transpower of New Zealand, Westbrook Power, Williams Energy Group, and Wisconsin Electric Power Company. The views presented here are not necessarily attributable to any of those mentioned, and any remaining errors are solely the responsibility of the author. (http://ksgwww.harvard.edu/people/whogan).

"ISOs," and finally, organizations for transmission loading relief. The different models present alternatives for the mix of responsibility of the necessary system operator. At one end of the spectrum, a Transco is an independent entity both owning the transmission assets and controlling system operations. By contrast, a Gridco is an entity owning the transmission assets but not responsible for system operations. System operations may be separated into a power exchange (PX) and transmission operations, or combined under an independent system operator (ISO). And finally, whatever regional choices are made, there must be institutions for coordinating transmission loading relief (TLR) across the regions.

The developing experiences around the world provide insight into the options and implications of alternative models. Comparison of these models also provides further information about some of the details of market operations. It is apparent from this experience that there must be a close connection between the design of options for market flexibility and the pricing principles for use of the transmission grid. If prices closely reflect operating conditions and marginal costs, then market participants can have numerous choices in the way they use the transmission system. However, if pricing does not conform to the operating conditions, then substantial operating restrictions must be imposed to preserve system reliability. Customer flexibility and choice require efficient pricing; inefficient pricing necessarily limits market flexibility.

Examples of failure and success illustrate the close connection between pricing provisions and operating rules. Events from markets as diverse as California, Australia, England and Wales, the New England Power Pool (NEPOOL), and the Pennsylvania-New Jersey-Maryland Interconnection (PJM) illuminate the basic point. The conclusion supports the integrated independent system operator (ISO), de jure on its own or de facto within a larger transmission organization, with locational marginal cost pricing rules, as the model most likely to be successful in supporting competitive markets with customer choices while preserving system reliability.

INSTITUTIONS AND MARKETS

A central problem in the development of competitive electricity markets arises from the need for a system operator who can manage the complex short-term interactions in the network and maintain system reliability. There is no choice. There must be a system operator. The only open questions are with the rules the system operator will apply and the governance of its activities.[2] The development of ISOs has proceeded steadily in the worldwide restructuring of electricity markets. There are significant advantages in this approach. Control of the use of the transmission grid means control of the dispatch, at least at the margin, because adjusting the dispatch is the principal (or, in some cases, only) means of affecting the flow of power on the grid. That this system operator should also be independent of the existing electric utilities and

---

[2] William W. Hogan, "Independent System Operator: Pricing and Flexibility in a Competitive Electricity Market," Center for Business and Government, Harvard University, February 1998.

other market participants is attractive in its simplicity in achieving equal treatment of all market participants. The ISO provides an essential service, but does not compete in the energy market.

There is a great deal of debate in the United States about the Federal Energy Regulatory Commission's (FERC) authority to mandate membership in an ISO, the need for such mandates, and the possibility that ISOs might be only a transitional arrangement. For example, there have been suggestions that ownership of the wires (a Gridco) combined with system operations (an SO) could produce an independent transmission company (a Transco) that would be different from an ISO, or an alternative that might be precluded by an ISO.

There is a continuing debate about the best model for organizing coordination and control of the transmission system.

Transco. An independent company that combines ownership of the grid and responsibility for system operations in managing the use of the grid. May be a for-profit or not-for-profit entity. (National Grid Company in England and Wales.)

Gridco. An independent company that owns the grid but does not have responsibility for operating the system. Works in conjunction with a system operator. May be a for-profit or not-for-profit entity. (GPU PowerNet in Victoria, Australia)

ISO/PX. An independent system operator with restrictions to allow for separate operation of a power exchange. (California ISO and PX.)

ISO. An independent system operator that has responsibility for managing use of the grid and coordinating the spot market. (Pennsylvania-New Jersey-Maryland Interconnection, PJM.)

TLR. The institution for coordinating transmission loading relief across regional system operators. (NERC Security coordinators in the U.S. Eastern Interconnect.)

A symptom of the confusion over the rules for a competitive market is in the parallel activities devoted to the discussion of ISOs, the FERC's OASIS system for transmission scheduling,[3] North American Electric Reliability Council (NERC) security coordinators for transmission line loading relief, and the FERC's earlier Capacity Reservation Tariff (CRT) proposals. Although these packages tend to be discussed in isolation, there is substantial overlap in that they all provide alternative approaches for the same core problem: rationing use of scarce

[3] Open Access Same time Information System (OASIS), FERC Order 889, Final Rule, Washington, DC, April 24, 1996.

transmission capacity. Furthermore, the approaches tend to be mutually inconsistent: some ISO models include bid-based economic dispatch; OASIS (in practice, if not in theory) is built around the flawed contract-path model; NERC's tagging rules and line loading relief procedures struggle to undo the contract path fiction, in order to deal with power flow realities and the commercial complications of administrative curtailments; the CRT would move all the way to a point-to-point reservation system with economic rationing.[4] We can't do all of these at the same time. And the attendant problems of coordinating trade across regions may be some of the most vexing for the regulators and the competitive market.

The unwelcome news for the regulators is that the hard problem of allocating scarce transmission capacity is made much more difficult by the move to competitive markets. In effect, we have taken the black box of the vertically integrated industry, opened it, and unbundled control of the various gears. In order for the system to work, however, the gears not only have to turn--they have to mesh. This is especially true in the very short-run, as we move closer to real-time operations.

Everyone wants non-discrimination and the maximum possible degree of flexibility for market participants. But to provide this flexibility, and make sure the gears mesh, it will be necessary to align the incentives of the participants with the success of the overall market. Either the incentives must match the system realities, or the pricing and access rules will be restrictive and dictate customer choices. Furthermore, the role of the system operator inevitably will encompass both reliability and commercial issues. The supposed distinction between reliability and economics is a mirage which will provide no comfort in practice. The nature of the electric grid dictates that decisions motivated by reliability concerns will have substantial commercial impacts, especially when the system is constrained and the decisions matter most. The only issue is the degree to which we will be explicit about the interaction between reliability and economics, in order to improve both efficiency and transparency.

The FERC refers to the range of models as regional transmission organizations, a term intended to encompass many models. Here we focus on the implications for competition in generation, and the rules for the wholesale market. With a focus on market institutions needed to support competitive markets, a critical summary of the debate over transmission models highlights the importance of system operations and real-time dispatch.

Transco

The Transco model as defined here emphasizes the combined responsibility for ownership of the wires and conduct of system operations. A Transco is a single regional entity which owns and operates the transmission system, but is independent of generation and load. The emphasis on control of system operations isolates one of the key elements that defines the relationship with the design of institutions for the competitive market.

---

[4] Michael Cadwalader, Scott Harvey, William Hogan, and Susan Pope, "Market Coordination of Transmission Loading Relief Across Multiple Regions," Center for Business and Government, Harvard University, December 1, 1998.

It seems only natural that ownership of an asset should imply control of its use. However, unlike most other markets, this link between ownership and control of operations is literally not possible for an interconnected electric network. Absent a single entity for the entire grid, there is no avoiding the necessity of have operations controlled at least in part by someone else. Hence, for electricity, setting the rules for how you use your own asset is unavoidable. The complications do not disappear through a simple change of ownership and governance.

The leading proposals call for regulated profit-making entities.[5] In part, the motivation for creating a Transco is to exploit the incentive effects of the profit motive. Presumably the profit opportunities would provide inducement for improved operations and market responsive investment. At a minimum, with ownership of significant assets, there is an argument that regulators would have greater leverage in controlling the performance of Transcos.

The strongest claims are that the profit motive is all that would be needed, and with incentive regulation the Transco could be left to devise its own rules for transmission access, operations and detailed pricing. By this argument, mere establishment of a for-profit Transco would dispense with the difficulties of evaluating the pricing and access rules for transmission and system operations.[6] Apparently through some type of incentive regulation, an independent Transco would support a non-discriminatory, competitive electricity market that meets the FERC's goals. While this may be a theoretical possibility, there is no known system of incentive regulation that could achieve this result. The difficulties to be overcome would begin with the same set of problems that complicate the process of setting the rules for system operation. At the core is the uncomfortable reality that there is no simple definition of the output of the transmission system. Efficient transmission is far more than electric throughput--it is a complex service with many dimensions and substantial network interactions. Were this not true, there would probably be no need for a system operator in the first place.

The Federal Trade Commission (FTC) has identified a flaw in the argument that a Transco would necessarily have the right incentives to support a competitive electricity market.[7] The central problem appears in the possible substitution between transmission and generation. We are experienced from many years of utility investment planning analysis that there is always a tradeoff between generation and transmission solutions when the system becomes constrained. It follows then that ownership of the wires and control of system operations (which means controlling the dispatch) would create an inherent conflict of interest for a Transco, with incentives to tilt operations to induce investment in transmission.

---

[5] Initial public announcements by Entergy, Northern States Power, and First Energy.

[6] For example, see Curt L. Hébert, Jr., "Moving the RTO Debate," The Electricity Journal, March 1999, pp. 20-23.

[7] Federal Trade Commission, Before Public Service Commission of the State of Mississippi, Docket No. 96-UA-389, August 28, 1998.

Furthermore, the assumption that it would be an easy matter to set the proper incentives for a Transco, incentives sufficient to leave to management the choice of rules and procedures for system operations, runs counter to the whole notion of electricity restructuring and greater reliance on the market. If we were so confident that we knew how to regulate such monopolies, then there would no need for restructuring and unbundling.[8] Quite to the contrary, it is a difficult matter to set such incentives.

An independent Transco may be attractive, and it could be the next stage or the end stage. But the very complexities that dictate the need for a system operator mean that it will not be an easy matter to structure the rules for system operations, nor would it be easy to structure incentives for a monopoly to discover the rules on its own. Providing appropriate incentives for the transmission system is a major difficulty in restructured electricity systems around the world. Some problems might be different under the different transmission models, such as the approach to providing incentives for grid maintenance and expansion,[9] but all the puzzles about the operating rules would appear again in this new guise. Hence, it is not likely that the Transco incentives could be developed so easily as to leave design of the system operation rules and pricing to the Transco monopoly alone. Somewhere in the company would be a system operator that must be "ringed fenced" from the rest of the corporation, to have its own independent rules and pricing structures that support the public interest in a competitive market, not only the private interests of the monopoly Transco.[10] The FERC will face the task of evaluating and approving the rules for pricing and access.[11] And this applies to the not-so-independent Transcos that are embedded in the vertically integrated utilities, as well as to new independent Transcos that might be divested from the utilities.

In the end, therefore, it is unlikely that the Transco would avoid any of the conceptual and design challenges that must be addressed in creating an ISO. In this sense, it would be a mistake to cast a Transco as an alternative to an ISO in any way other than the formalities of governance. The same questions that have appeared in the specification of the rules for the market, for access and pricing, would appear in establishing the rules for the Transco. The main problems cannot be avoided. In effect, the result is likely to be a de facto ISO within the de jure corporate structure of a Transco. Or we could think of a Transco as an ISO that acquires

---

[8] Lawrence J. Spiwak, "You Say ISO, I Say Transco, Let's Call the Whole Thing Off," Public Utilities Fortnightly, March 15, 1999.

[9] William W. Hogan, "Transmission Investment and Competitive Electricity Markets," Center for Business and Government, Harvard University, April 1998.

[10] Fiona Woolf, Cameron McKenna, comments on Panel 3, "Regulation, Governance, and Independence," FERC Public Conference Concerning the Commission's Policy on Independent System Operators, April 16, 1998. Examples with careful attention to the market rules by which the operators within the Transco pursue dispatch and pricing can be found in the National Grid Company of England and Wales or Transpower in New Zealand.

[11] William L. Massey, "Policy on Regional Transmission Organizations: Five Pitfalls FERC Must Avoid," The Electricity Journal, March 1999, pp. 13-19.

ownership of the wires.

If we are not relying on the profit incentive alone to produce the rules for system operations, then other approaches to the Transco model might capture some of the benefits of better coordination of transmission investment and wires maintenance combined with an understanding of the needs of system operations. Here the large public power authorities in the United States provide an alternative model with non-profit organizations.[12] This is an old debate, with strong views and conflicting evidence. In the choice between the for-profit and the not-for-profit model, it may be the other details on regional coverage, legal restrictions on the transition, and the model for market operations would be more decisive.

Finally, a major hurdle for the widespread embrace of the Transco model in a large country like the United States would be in creating transmission companies that match the regional requirements of system operations. This is easy in New Zealand where there is already a single transmission owner. It would be much more difficult in the United States, Japan, or Europe with their large interconnected systems. The combination of system operations and ownership of only some of the wires might be much more problematic from the perspective of the owners of the other interconnected wires. Either these other transmission owners must surrender control of operations, foregoing all the presumed benefits of the Transco model, or operations must be balkanized to follow the pattern of ownership. If it is not an easy matter to change the patterns of ownership of the wires, therefore, reliance on the Transco model would substantially hinder the development of integrated markets with broad regional coverage.

Gridco

A Gridco is a regional entity that owns transmission wires and is independent of generation and load. The Gridco is not responsible for controlling use of the system, and must be paired with a system operator. Many of the advantages of the Transco model would apply to the Gridco approach, but without all of the problems.

Control of operations by an ISO is compatible with the Gridco model. The rules for access and pricing would be the same as under the regime where traditional utilities own the grid. The distinction of the Gridco is that maintenance and expansion of the grid could be the responsibility of the Gridco, which is also independent of generation and load.

As with Transcos, the leading proposals call for regulated for-profit entities, such as the strategies embodied in the public announcements of New England Electric System in New England or General Public Utilities in Pennsylvania-New Jersey. In both these cases, the wires company is separated from ownership of generation and from system operations. However, in these cases the companies still own distribution wires, and are not strictly transmission companies as understood in the pure Gridco model. Sharper examples of a pure Gridco approach are found in the transmission companies in Australia such as GPU PowerNet in Victoria, which owns and

---

[12] Frank McCarmant, Vincent Tobin, and Stephen Pelcher, "Uncrossing the Wires: Transmission in a Restructured Market," The Electricity Journal, March 1999, pp. 24-35.

maintains the transmission wires, but does not own distribution systems and leaves system operations to the independent system operator in the National Electricity Market Management Company (NEMMCO).

The arguments of the large public power authorities apply as well for a mix that includes non-profit Gridcos. The incentives for the Gridco, which would own significant assets, would be similar to those of the Transco, but without the conflicts of interest in operations identified by the FTC.

Because of the separation from operations, regional coverage of the Gridco ownership of the wires need not and probably would not coincide with the regional coverage of system operations. This would be a great simplification compared to the Transco model. It would allow an evolution of Gridcos, with different models, without confronting the complications of balkanized operations. The developing Gridcos would be able to manage transmission investment and maintenance, with appropriate incentive regulation. Providing the proper incentives would not be easy, but it would be possible to pursue market incentives with fewer difficulties in isolating the control of system operations.

Power Exchanges

The debate launched in the California restructuring spawned separate institutions for the operation of the spot market through a power exchange (PX) and control of system operations through an ISO. Here the independent system operator functions in conjunction with a separate and distinct power exchange responsible for market operations, with separate rules and pricing for each. In this case, neither the ISO nor PX owns transmission lines.

The viability of the distinction between the functions of market operations and system operations depend on the time horizon and the relative importance of network interactions. For the short-run, the two functions are difficult (impossible) to separate.[13] Over the short-run, maintaining a distinction between the ISO and the PX requires creation of complex rules to restrict the system operator. It is well recognized that if the system operator performs its functions through use of a voluntary, bid-based, security-constrained, economic dispatch--following the principles power systems have used for decades--the separate power exchange would have little to do other than arrange accounting settlements. Hence, in California, where we find the only model with this formal attempt to separate the spot market from system operations, the design precludes the ISO from pursuing an economic dispatch and segments interdependent functions, reducing options and increasing costs. As summarized below, the results with this approach have been problematic, at best, and a number of initiatives are

[13] W. Hogan, "A Wholesale Pool Spot Market Must Be Administered by the Independent System Operator: Avoiding the Separation Fallacy," The Electricity Journal, December 1995, pp. 26-37.

underway that will, in effect, undo the artificial separation of markets.[14] Restrictions on ISOs reappear in various proposals elsewhere that limit the use of economic dispatch and transmission coordination, assuming that the complex interactions can somehow be internalized in a market, even without a formal power exchange.[15] Inevitably these approaches reduce transmission capacity, socialize costs and add to the complexity of real operations.

Over horizons where network interaction might not be as important, the advantages for integration of the power exchange and system operations would be reduced. For example, in Norway there is a market which functions as a power exchange separate from the ISO, for trading of contracts and establishment of schedules. But in the end, the true real-time spot market in Norway is the final regulation market administered by the ISO.

Independent System Operator

The independent system operator provides a dispatch function that coordinates the spot market. As discussed below, the basic ISO model has been extensively developed in its various incarnations around the world.[16] The ISO does not own transmission lines. If there is a separate entity called a Power Exchange, it does not have responsibility for coordinating the spot market and transmission usage. The PX may handle bidding and settlements, such as with Electricity Market Company (EMCO) in New Zealand. But in New Zealand the real-time dispatch implementation falls to the part of Transpower that is the de facto ISO. In many cases, there is no separate PX with any special status, as for example in the Pennsylvania-New Jersey-Maryland Interconnection (PJM), the New England ISO, the Australian NEMMCO, and so on.

The services provided by the ISO are complex and interconnected. It is a challenge to find the best mix of unbundled activities and associated pricing rules. The key is to match the degree of customer choice with the pricing incentives. Where customers have flexibility, such as between spot market transactions and bilateral transmission scheduling, it is important to get the prices right. There are many models, each with its own nuances. But there is a core model built on the basic economics of electricity systems, as discussed below.

The appropriate size and regional coverage of the ISO depends on many factors, including the degree of coordination required across the entities in arranging for transmission loading relief. The ISO model is fully compatible with the creation of independent Gridcos, and enjoys the advantage over a Transco that the regional coverage of the single ISO does not have to match that of possibly multiple Gridcos.

---

[14] For example, see Frank Wolak, Robert Nordhaus, and Carl Shapiro, "Report on the Redesign of Markets for Ancillary Services and Real-Time Energy," Market Surveillance Committee of the California Independent System Operator, March 25, 1999.

[15] For example, this is the proposal for the Midwest Independent System Operator.

[16] William W. Hogan, "Independent System Operator: Pricing and Flexibility in a Competitive Electricity Market," Harvard-Japan Project on Energy and the Environment, Harvard University, March 1998.

Transmission Loading Relief

In large interconnected grids with multiple areas under separate control, regional system operators must coordinate use of the transmission grid. Transmission loading relief (TLR) is required when system constraints would be violated. The rules for inter-regional coordination interact strongly with the pricing and access rules within the regions. The experience in the United States has been that there has been too little in the way of coordination of the TLR rules with the requirements and expectations of the developing markets.

In the United States, the North American Electric Reliability Council (NERC) filled the vacuum in developing a mechanism and institutional framework for TLR. However, the institutional design limits imposed or assumed by NERC required non-market mechanisms for curtailing transactions. In effect, the NERC approach embraced the fallacy of the separation of markets and reliability, assuming that it would be possible to have reliability rules that would be either unimportant or neutral in their commercial effects. The resulting TLR system was cumbersome, reduced real transmission capacity, and had severe impacts on the market, contributing to problems in the Mid-West during the summer of 1998 that produced $7000/MWh transactions.[17]

With TLR integrated in the market, prices and bids would matter. The FERC directed NERC to develop more economic systems. There are alternative market mechanisms available in principle. For example, the PJM system proposed implementing the first consistent market mechanism for managing TLR by allowing participants to choose to pay for congestion. This is a separate topic still under active consideration in the United States. The lesson is that the TLR rules must be developed to be consistent with the institutions of the electricity market. The reliability driven concerns for use of the transmission system cannot be separated from system operations or from the activities of the market, at least in the short-run. Network interactions are strong, and the same forces that create the need for an ISO drive the need for market driven methods of adjusting use of the transmission system.

ECONOMICS OF A COMPETITIVE ELECTRICITY MARKET

A general framework that encompassed the essential economics of electricity markets provides a point of reference for evaluating market design elements.[18] Here we focus on

---

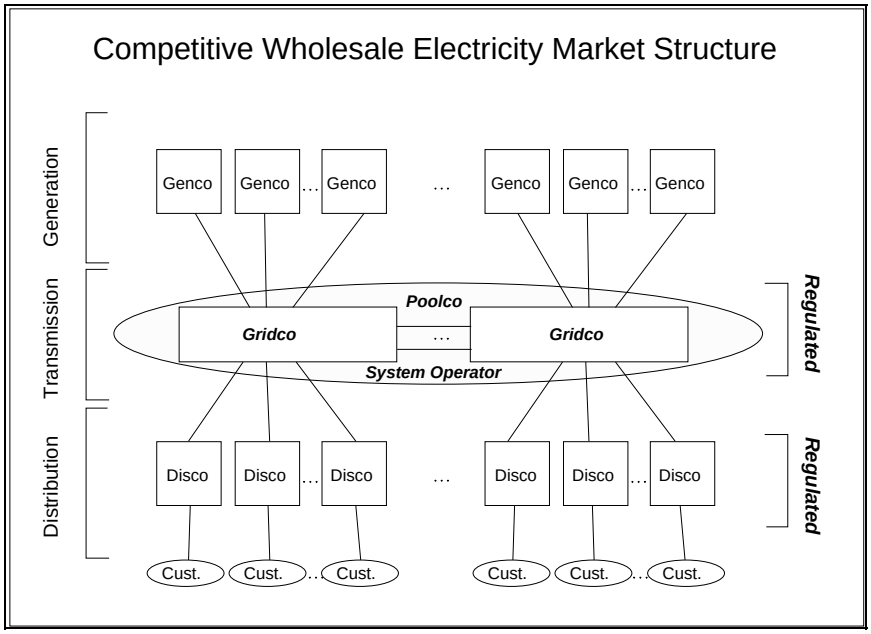
[17] Michael Cadwalader, Scott Harvey, William Hogan, and Susan Pope, "Market Coordination of Transmission Loading Relief Across Multiple Regions," Center for Business and Government, Harvard University, December 1, 1998.

[18] This summary comes primarily from William W. Hogan, "Transmission Investment and Competitive Electricity Markets," Center for Business and Government, Harvard University, April 1998. The issues are developed further there, but summarized here for completeness given the central importance of the basic economics in the case of electricity. See also, William W. Hogan, "Competitive Electricity Markets: A Wholesale Primer," Center for Business and Government, Harvard University, December 1998.

theimplications for competition in generation, and the rules for the wholesale market. The treatment of competition for other contestable elements, such as retail services, is important but need not affect the design of the wholesale market. This framework provides a background for evaluating the prescriptions for ISOs and related market institutions.

Competitive Market Design

Reliable operation is a central requirement and constraint for any electricity system. Given the strong and complex interactions in electric networks, current technology with a free-flowing transmission grid dictates the need for a system operator that coordinates use of the transmission system. Control of transmission usage means control of dispatch, which is the principal or only means of adjusting the use of the network. Hence, open access to the transmission grid means open access to the dispatch as well. In the analysis of electricity markets, therefore, a key focus is the design of the interaction between transmission and dispatch, both procedures and pricing, to support a competitive market.



To provide an overview of the operation of an efficient, competitive wholesale electricity market, it is natural to distinguish between the short-run operations coordinated by the system operator and long-run decisions that include investment and contracting. Market participants are price takers and include the generators and eligible customers. For this discussion, distributors are included as customers in the wholesale market, operating at arm's length from generators. The system is much simpler in the very short run when it is possible to give meaningful definition to concepts such as opportunity cost. Once the short-run economics are established, the long-run requirements become more transparent. Close attention to the connection between short- and long-run decisions isolates the special features of the electricity market.

Short-Run Market

The short run is a long time on the electrical scale, but short on human scale – say, half an hour. The short-run market is relatively simple. In the short run, locational investment decisions have been made. Power plants, the transmission grid, and distribution lines are all in place. Customers and generators are connected and the work of buyers, sellers, brokers and other

service entities is largely complete. The only decisions that remain are for delivery of power, which in the short-run is truly a commodity product.

On the electrical scale, much can happen in half an hour and the services provided by the system include many details of dynamic frequency control and emergency response to contingencies. Due to transaction costs, if nothing else, it would be inefficient to unbundle all of these services, and many are covered as average costs in the overhead of the system. How far unbundling should go is an empirical question. For example, real power should be identified and its marginal cost recognized, but should this extend to reactive power and voltage control as well? Or to spinning reserve required for emergency supplies? For the sake of the present discussion, focus on real power and assume that further unbundling would go beyond the point of diminishing returns in the short-run market.

Over the half hour, the market operates competitively to move real power from generators to customers. Generators have a marginal cost of generating real power from each plant, and customers have different quantities of demand depending on the price at that half hour. The collection of generator costs stacks up to define the generation "merit order," from least to most expensive. This merit order defines the short-run marginal-cost curve, which governs power supply. Similarly, customers have demands that are sensitive to price, and higher prices produce lower demands. Generators and customers do not act unilaterally; they provide information to the dispatcher to be used in a decision process that will determine which plants will run at any given half hour. Power pools provide the model for achieving the most efficient dispatch given the short-run marginal costs of power supply. Although dispatchable demand is not always included, there is nothing conceptually or technically difficult about this extension. The system operator controls operation of the system to achieve the efficient match of supply and demand.

This efficient central dispatch can be made compatible with the market outcome. The fundamental principle is that for the same load, the least-cost dispatch and the competitive-market dispatch are the same. The principal difference between the traditional power pool and the market solution is the price charged to the customer. In the traditional power pool model, customers pay and generators receive average cost, at least on average. Marginal cost implicitly determines the least-cost dispatch, and marginal cost is the standard determinant of competitive market pricing.

An important distinction between the traditional central dispatch and the decentralized market view is found in the source of the marginal-cost information for the generator supply curve. Traditionally the cost data come from engineering estimates of the energy cost of generating power from a given plant at a given time. However, relying on these engineering estimates is problematic in the market model since the true opportunity costs may include other features, such as the different levels of maintenance, that would not be captured in the fuel cost. Replacement of the generator’s engineering estimates (that report only incremental fuel cost) with the generator’s market bids is the natural alternative. Each bid defines the minimum acceptable price that the generator would accept to run the plant in the given half hour. And these bids serve as the guide for the dispatch.

As long as the generator receives the market clearing price, and there are enough competitors so that each generator assumes that it will not be providing the marginal plant, then the optimal bid for each generator is the true marginal cost: To bid more would only lessen the chance of being dispatched, but not change the price received. To bid less would create the risk of running and being paid less than the cost of generation for that plant. Hence, with enough competitors and no collusion, the short-run central dispatch market model can elicit bids from buyers and sellers. The system operator can treat these bids as the supply and demand and determine the balance that maximizes benefits for producers and consumers at the market equilibrium price. Hence, in the short run electricity is a commodity, freely flowing into the transmission grid from selected generators and out of the grid to the willing customers. Every half hour, customers pay and generators receive the short-run marginal-cost (SRMC) price for the total quantity of energy supplied in that half hour. Everyone pays or receives the true opportunity cost in the short run. Payments follow in a simple settlement process.

Transmission Congestion

This overview of the short-run market model is by now familiar and found in operation in many countries. However, this introductory overview conceals a critical detail that would be relevant for transmission pricing. Not all power is generated and consumed at the same location. In reality, generating plants and customers are connected through a free-flowing grid of transmission and distribution lines.

In the short-run, transmission too is relatively simple. The grid has been built and everyone is connected with no more than certain engineering requirements to meet minimum technical standards. In this short-run world, transmission reduces to nothing more than putting power into one part of the grid and taking it out at another. Power flow is determined by physical laws, but a focus on the flows – whether on a fictional contract path or on more elaborate allocation methods – is a distraction. The simpler model of input somewhere and output somewhere else captures the necessary reality. In this simple model, transmission complicates the short-run market through the introduction of losses and possible congestion costs.

Transmission of power over wires encounters resistance, and resistance creates losses. Hence the marginal cost of delivering power to different locations differs at least by the marginal effect on losses in the system. Incorporating these losses does not require a major change in the theory or practice of competitive market implementation. Economic dispatch would take account of losses, and the market equilibrium price could be adjusted accordingly. Technically this would yield different marginal costs and different prices, depending on location, but the basic market model and its operation in the short-run would be preserved.

Transmission congestion has a related effect. Limitations in the transmission grid in the short run may constrain long-distance movement of power and thereby impose a higher marginal cost in certain locations. Power will flow over the transmission line from the low cost to the high cost location. If this line has a limit, then in periods of high demand not all the power that could be generated in the low cost region could be used, and some of the cheap plants would be "constrained off." In this case, the demand would be met by higher cost plants that

absent the constraint would not run, but due to transmission congestion would be then "constrained on." The marginal cost in the two locations differs because of transmission congestion. The marginal cost of power at the low cost location is no greater than the cost of the cheapest constrained-off plant; otherwise the plant would run. Similarly, the marginal cost at the high cost location is no less than the cost of the most expensive constrained-on plant; otherwise the plant would not be in use. The difference between these two costs, net of marginal losses, is the congestion rental.

This congested-induced marginal-cost difference can be as large as the cost of the generation in the unconstrained case. If a cheap coal plant is constrained off and an oil plant, which costs more than twice as much to run, is constrained on, the difference in marginal costs by region is greater than the cost of energy at the coal plant. This result does not depend in any way on the use of a simple case with a single line and two locations. In a real network the interactions are more complicated – with loop flow and multiple contingencies confronting thermal limits on lines or voltage limits on buses – but the result is the same. It is easy to construct examples where congestion in the transmission grid leads to marginal costs that differ by more than 100% across different locations.

If there is transmission congestion, therefore, the short-run market model and determination of marginal costs must include the effects of the constraints. This extension presents no difficulty in principle. The only impact is that the market now includes a set of prices, one for each location. Economic dispatch would still be the least-cost equilibrium subject to the security constraints. Generators would still bid as before, with the bid understood to be the minimum acceptable price at their location. Customers would bid also, with dispatchable demand and the bid setting the maximum price that would be paid at the customer's location. The security-constrained economic dispatch process would produce the corresponding prices at each location, incorporating the combined effect of generation, losses and congestion. In terms of their own supply and demand, everyone would see a single price, which is the SRMC price of power at their location. If a transmission price is necessary, the natural definition of transmission is supplying power at one location and using it at another. The corresponding transmission price would be the difference between the prices at the two locations.

This same framework lends itself easily to accounting extensions to explicitly include bilateral transactions. The bilateral schedules would be provided to the system operator. Those not scheduled would bid into the pool-based spot market. This is often described as the "residual pool" approach. For market participants who wish to schedule transmission between two locations, the opportunity cost of the transmission is just this transmission price of the difference between spot prices at the two locations. This short-run transmission usage pricing, therefore, is efficient and non-discriminatory. In addition, the same principles could apply in a multi-settlement framework, with day-ahead scheduling and real-time dispatch. These extensions could be important in practice, but would not fundamentally change the outline of the structure of electricity markets.

This short-run competitive market with bidding and centralized dispatch is consistent

with economic dispatch. The locational prices define the true and full opportunity cost in the short run. Each generator and each customer sees a single price for the half hour, and the prices vary over half hours to reflect changing supply and demand conditions. All the complexities of the power supply grid and network interactions are subsumed under the economic dispatch and calculation of the locational SRMC prices. These are the only prices needed, and payments for short-term energy are the only payments operating in the short run, with administrative overhead covered by rents on losses or, if necessary, a negligible markup applied to all power. The system operator coordinates the dispatch and provides the information for settlement payments, with regulatory oversight to guarantee comparable service through open access to the pool run by the system operator through a bid-based economic dispatch.

With efficient pricing, users have the incentive to respond to the requirements of reliable operation. Absent such price incentives, choice would need to be curtailed and the market limited, in order to give the system operator enough control to counteract the perverse incentives that would be created by prices that did not reflect the marginal costs of dispatch. A competitive market with choice and customer flexibility depends on getting the usage pricing right.

Long-Run Market Contracts

With changing supply and demand conditions, generators and customers will see fluctuations in short-run prices. When demand is high, more expensive generation will be employed, raising the equilibrium market prices. When transmission constraints bind, congestion costs will change prices at different locations.

Even without transmission congestion constraints, the spot market price can be volatile. This volatility in prices presents its own risks for both generators and customers, and there will be a natural interest in long-term mechanisms to mitigate or share this risk. The choice in a market is for long-term contracts.

Traditionally, and in many other markets, the notion of a long-term contract carries with it the assumption that customers and generators can make an agreement to trade a certain amount of power at a certain price. The implicit assumption is that a specific generator will run to satisfy the demand of a specific customer. To the extent that the customer's needs change, the customer might sell the contract in a secondary market, and so too for the generator. Efficient operation of the secondary market would guarantee equilibrium and everyone would face the true opportunity cost at the margin.

However, this notion of specific performance stands at odds with the operation of the short-run market for electricity. To achieve an efficient economic dispatch in the short-run, the dispatcher must have freedom in responding to the bids to decide which plants run and which are idle, independent of the provisions of long-term contracts. And with the complex network interactions, it is impossible to identify which generator is serving which customer. All generation is providing power into the grid, and all customers are taking power out of the grid. In a competitive market, it is not even in the interest of the generators or the customers to restrict

their dispatch and forego the benefits of the most economic use of the available generation. The short-term dispatch decisions by the system operator are made independent of and without any recognition of any long-term contracts. In this way, electricity is not like other commodities.

This dictate of the physical laws governing power flow on the transmission grid does not preclude long-term contracts, but it does change the essential character of the contracts. Rather than controlling the dispatch and the short-run market, long-term contracts focus on the problem of price volatility and provide a price hedge not by managing the flow of power but by managing the flow of money. The short-run prices provide the right incentives for generation and consumption, but create a need to hedge the price changes. Recognizing the operation of the short-run market, there is an economic equivalent of the long-run contract for power that does not require any specific plant to run for any specific customer.

Consider the case first of no transmission congestion. In this circumstance, except for the small effect of losses, it is possible to treat all production and consumption as at the same location. Here the natural arrangement is to contract for differences against the equilibrium price in the market. A customer and a generator agree on an average price for a fixed quantity, say 100 MW at five cents. On the half hour, if the spot price is six cents, the customer buys power from the pool at six cents and the generators sells power for six cents. Under the contract, the generator owes the customer one cent for each of the 100 MW over the half hour. In the reverse case, with the pool price at three cents, the customer pays three cents to the pool, which in turn pays three cents to the generator, but now the customer owes the generator two cents for each of the 100 MW over the half hour.

In effect, the generator and the customer have a long-term contract for 100 MW at five cents. The contract requires no direct interaction with system operator other than for the continuing short-run market transactions. But through the interaction with system operator, the situation is even better than with a long-run contract between a specific generator and a specific customer. For now if the customer demand is above or below 100 MW, there is a ready and an automatic secondary market, namely the pool, where extra power is purchased or sold at the pool price. Similarly for the generator, there is an automatic market for surplus power or backup supplies without the cost and problems of a large number of repeated short-run bilateral negotiations with other generators. And if the customer really consumes 100 MW, and the generator really produces the 100 MW, the economics guarantee that the average price is still five cents. Furthermore, with the contract fixed at 100 MW, rather than the amount actually produced or consumed, the long-run average price is guaranteed without disturbing any of the short-run incentives at the margin. Hence the long-run contract is compatible with the short-run market.

The price of the generation contract would depend on the agreed reference price and other terms and conditions. Generators and customers might agree on dead zones, different up-side and down-side price commitments, or anything else that could be negotiated in a free market to reflect the circumstances and risk preferences of the parties. Whether generators pay customers, or the reverse, depends on the terms. However, system operator need take no notice of the contracts, and have no knowledge of the terms.

In the presence of transmission congestion, the generation contract is necessary but not sufficient to provide the necessary long-term price hedge. A bilateral arrangement between a customer and a generator can capture the effect of aggregate movements in the market, when the single market price is up or the single market price is down. However, transmission congestion can produce significant movements in price that are different depending on location. If the customer is located far from the generator, transmission congestion might confront the customer with a high locational price and leave the generator with a low locational price. Now the generator alone cannot provide the natural back-to-back hedge on fluctuations of the short-run market price. Something more is needed.

Transmission congestion in the short-run market raises another related and significant matter for the system operator. In the presence of congestion, revenues collected from customers will substantially exceed the payments to generators. The difference is the congestion rent that accrues because of constraints in the transmission grid. At a minimum, this congestion rent revenue itself will be a highly volatile source of payment to the system operator. At worse, if the system operator keeps the congestion revenue, incentives arise to manipulate dispatch and prevent grid expansion in order to generate even greater congestion rentals. System operation is a natural monopoly and the operator could distort both dispatch and expansion. If the system operator retains the benefits from congestion rentals, this incentive would work contrary to the goal of an efficient, competitive electricity market.

The convenient solution to both problems – providing a price hedge against locational congestion differentials and removing the adverse incentive for system operator – is to re-distribute the congestion revenue through a system of long-run transmission congestion contracts operating in parallel with the long-run generation contracts. Just as with generation, it is not possible to operate an efficient short-run market that includes transmission of specific power to specific customers. However, just as with generation, it is possible to arrange a transmission congestion contract that provides compensation for differences in prices, in this case for differences in the congestion costs between different locations across the network.

The transmission congestion contract for compensation would exist for a particular quantity between two locations. The generator in the example above might obtain a transmission congestion contract for 100 MW between the generator's location and the customer's location. The right provide by the contract would not be for specific movement of power but rather for payment of the congestion rental. Hence, if a transmission constraint caused prices to rise to six cents at the customer's location, but remain at five cents at the generator's location, the one cent difference would be the congestion rental. The customer would pay the pool six cents for the power. The pool would in turn pay the generator five cents for the power supplied in the short-run market. As the holder of the transmission congestion contract, the generator would receive one cent for each of the 100 MW covered under the transmission congestion contract. This revenue would allow the generator to pay the difference under the generation contract so that the net cost to the customer is five cents as agreed in the bilateral power contract. Without the transmission congestion contract, the generator would have no revenue to compensate the customer for the difference in the prices at their two locations. The transmission congestion

contract completes the package.

When only the single generator and customer are involved, this sequence of exchanges under the two types of contracts may seem unnecessary. However, in a real network with many participants, the process is far less obvious. There will be many possible transmission combinations between different locations. There is no single definition of transmission grid capacity, and it is only meaningful to ask if the configuration of allocated transmission flows is feasible. However, the net result would be the same. Short-run incentives at the margin follow the incentives of short-run opportunity costs, and long-run contracts operate to provide price hedges against specific quantities. The system operator coordinates the short-run market to provide economic dispatch. The system operator collects and pays according to the short-run marginal price at each location, and the system operator distributes the congestion rentals to the holders of transmission congestion contracts. Generators and customers make separate bilateral arrangements for generation contracts. Unlike with the generation contracts, the system operator's participation in coordinating administration of the transmission congestion contracts is necessary because of the network interactions, which make it impossible to link specific customers paying congestion costs with specific customer receiving congestion compensation. If a simple feasibility test is imposed on the transmission congestion contracts awarded to customers, the aggregate congestion payments received by the system operator will fund the congestion payment obligations under the transmission congestion contracts. Still, the congestion prices paid and received will be highly variable and load dependent. Only the system operator will have the necessary information to determine these changing prices, but the information will be readily available embedded in all the pool's locational prices. The transmission congestion contracts define payment obligations that guarantee protection from changes in the congestion rentals.

The transmission congestion contract can be recognized as equivalent to an advantageous form of "physical" transmission right. Were it possible to define usage of the transmission system in terms of physical rights, it would be desirable that these rights have two features. First, they could not be withheld from the market to prevent others from using the transmission grid. Second, they would be perfectly tradable in a secondary market that would support full reconfiguration of the patterns of network use at no transaction cost. This is impossible with any known system of transmission rights that parcel up the transmission grid. However, in a competitive electricity market with a bid-based, security-constrained economic dispatch, transmission congestion contracts are equivalent to just such perfectly tradable transmission rights. Hence we can describe transmission congestion contracts either as financial contracts for congestion rents or as perfectly tradable physical transmission rights.

If the transmission congestion contracts have been fully allocated, then the system operator will be simply a conduit for the distribution of the congestion rentals. The operator would no longer have an incentive to increase congestion rentals: any increase in congestion payments would flow only to the holders of the transmission congestion contracts. The problem of supervising the dispatch monopoly would be greatly reduced. And through a combination of generation contracts and transmission congestion contracts, participants in the electricity market can arrange price hedges that could provide the economic equivalent of a long-term contract for

specific power delivered to a specific customer.

Further to the application of these ideas, locational marginal cost pricing lends itself to a natural decomposition. For example, even with loops in a network, market information could be transformed easily into a hub-and-spoke framework with locational price differences on a spoke defining the cost of moving to and from the local hub, and then between hubs. This would simplify without distorting the locational prices. A contract network could develop that would be different from the real network without affecting the meaning or interpretation of the locational prices.

With the market hubs, the participants would see the simplification of having a few hubs that capture most of the price differences of long-distance transmission. Contracts could develop relative to the hubs. The rest of the sometimes important difference in locational prices would appear in the cost of moving power to and from the local hub. Commercial connections in the network could follow a configuration convenient for contracting and trading. The separation of physical and financial flows would allow this flexibility.

The creation or elimination of hubs would require no intervention by regulators or the system operator. New hubs could arise as the market requires, or disappear when not important. A hub is simply a special node within a zone. The system operator still would work with the locational prices, but the market would decide on the degree of simplification needed. However, everyone would still be responsible for the opportunity cost of moving power to and from the local hub. There would be locational prices and this would avoid the substantial incentive problems of averaging prices.

Long-Term Market Investment

Within the contract environment of the competitive electricity market, new investment occurs principally in generating plants, customer facilities and transmission expansions. In each case, corresponding contract-right opportunities appear that can be used to hedge the price uncertainty inherent in the operation of the competitive short-run market.

In the case of investment in new generating plants or consuming facilities, the process is straightforward. Under the competitive assumption, no single generator or customer is a large part of the market, there are no significant economies of scale, and there are no barriers to entry. Generators or customers can connect to the transmission grid at any point subject only to technical requirements defining the physical standards for hookup. If they choose, new customers or new generators have the option of relying solely on the short-run market, buying and selling power at the locational price determined as part of the half-hourly dispatch. The system operator makes no guarantees as to the price at the location. The system operator only guarantees open access to the pool at a price consistent with the equilibrium market. The investor takes all the business risk of generating or consuming power at an acceptable price.

If the generator or customer wants price certainty, then new generation contracts can be struck between a willing buyer and a willing seller. The complexity and reach of these

contracts would be limited only by the needs of the market. Typically we expect a new generator to look for a customer who wants a price hedge, and for the generators to defer investing in new plant until sufficient long-term contracts with customers can be arranged to cover a sufficient portion of the required investment. The generation contracts could be with one or more customers and might involve a mix of fixed charges coupled with the obligations to compensate for price differences relative to the spot-market price. But the customer and generator would ultimately buy and sell power at their location at the half-hourly price.

If either party expects significant transmission congestion, then a transmission congestion contract would be indicated. If transmission congestion contracts are for sale between the two points, then a contract can be obtained from the holder(s) of existing rights. Or new investment can create new capacity that would support additional transmission congestion contracts. The system operator would participate in the process only to verify that the newly created transmission congestion contracts would be feasible and consistent with the obligation to preserve any existing set of transmission congestion contracts on the existing grid. Unlike the ambiguity in the traditional definition of transmission transfer capacity, there is a direct test to determine the feasibility of any new set of transmission congestion contracts for compensation – while protecting the existing rights – and the test is independent of the actual loads that may develop. Hence, incremental investments in the grid would be possible anywhere without requiring that everyone connected to the grid participate in the negotiations or agree to the allocation of the new transmission congestion contracts.

This happy resolution of the puzzle of transmission expansion and pricing through voluntary market forces alone is subject to at least two other important caveats. First, there still may be market failures even with the definition of a workable set of equivalent property rights. For example, with many small market participants, each benefiting a little from a large transmission investment, the temptation to free-ride on the economies of scale and scope may create a kind of prisoner's dilemma. Everyone would be better off sharing in the investment, but the temptation to free ride and avoid paying for the expense may overcome any ability to form a consortium or negotiate a contract. It may be that the investment could not go forward in a timely manner, at the right scale, or at all, without some regulatory entity that can mandate payment of the costs.[19] In this case, however, the task should be simplified by the ability to simultaneously allocate the benefits in the form of a share of the transmission congestion contracts. The market could take care of many, perhaps most, investments, and the regulatory option would be easier to implement when needed.

Second, operation of voluntary market forces would have little sway in the allocation of the costs for an existing transmission grid that already provides open access. The costs are sunk, and typically the sunk costs of the wires exceed the transmission congestion opportunity costs of using the grid. This is due, in large part, to the effects of the economies of scale.

---

[19] This situation appears to be what is described often as investments for reliability. However, with price responsive demand and security constrained economic dispatch, there is in principle no difference in reliability. The only difference created by the investment would be in the economic benefits of the actual dispatch.

Hence, given the choice of paying the sunk costs but avoiding the congestion costs, versus avoiding the sunk costs while using the system and paying the continuing cost of congestion, most users would prefer the latter. If the sunk costs are to be recovered in prospective payments, therefore, there must be some form of requirement to pay these costs as a condition for using the grid. The resulting access charges would be the functional equivalent of the contract payments for new investment.

PRESCRIPTIONS FOR (I)SOs

There are a few essential services related to coordinating use of the transmission grid where the ISO is both necessary and would have a significant comparative advantage. For example, although it is possible to design an ISO that ignores market preferences, the ISO would have a significant advantage in conducting its short-term coordination activities through an open spot market. In addition to the immediate efficiency improvements, the transparency and ease of entry for small participants would provide a wealth of benefits for promoting the long-run competitiveness of the market. These benefits would include a practical framework for implementing transmission rights, as embodied in the PJM approach.[20]

The PJM model and similar systems such as for New York, New Zealand, and so on, provide open access with non-discriminatory pricing. The critical short-run matter of congestion pricing and allocation of scarce transmission capacity through locational marginal cost pricing complements other components to deal with the longer-term issues that go beyond system operations. Transmission fixed costs are recovered primarily through system-wide (but not necessarily uniform) network service charges. The ISO administers both a spot market and bilateral schedules, while maintaining reliability under principles of bid-based, economic, security-constrained dispatch. Fixed transmission rights (FTRs, i.e., transmission congestion contracts) are available for congestion costs between locations, creating the market equivalent of perfectly tradeable physical transmission rights, providing tradable open-access transmission reservations in the only way that is likely to be feasible.

Market Design Pitfalls

There is little public policy justification for approving ISO rules that go in the direction of adding further complications while restricting participant choices. Hence, it would be appropriate to prescribe rules for the ISO that would support a competitive market in the real world, rather than an idealized world where network complications could be ignored. At a minimum, governments and regulators interested in market design should look with great skepticism on proposals that begrudgingly acknowledge that a certain function must be performed by the system operator, but then require that it be performed badly. For example:

---

[20] Scott M. Harvey, William W. Hogan, and Susan L. Pope, "Transmission Capacity Reservations and Transmission Congestion Contracts," Center for Business and Government, Harvard University, June 6, 1996, (Revised March 8, 1997).

Balancing Services and Penalties. The ISO must provide real time balancing to maintain system integrity. Balancing imposes costs, and those relying on the balancing services should pay these costs. However, a strong burden of proof should face those who would charge balancing penalties in excess of costs, or restrict voluntary access to balancing services.

Balancing Constraints. The ISO must maintain aggregate energy balance in the system, but there is no physical necessity and no public policy interest in requiring particular combinations of transactions to remain balanced. Quite the contrary. Individual balancing requirements both complicate the task for the ISO and provide a device to reinforce market power. This goes against the public interest.

Least-Cost (Re)Dispatch. The ISO must be able to (re)dispatch plants in order to manage transmission congestion. Rules designed to prevent the ISO from applying the familiar principles of economic dispatch run contrary to the notion of competitive markets and the public interest.

Voluntary Bidding. When doing an economic dispatch, it seems logical for the ISO to make the adjustments taking into account the preferences of the market participants as expressed by their voluntary bids. There should be a strong burden of proof for those who argue that it is necessary to restrict the voluntary bids, or discard consideration of some bids.

Transmission Rights and Dispatch. The ISO must coordinate the use of the transmission system. And once the actual use of the transmission system is determined, so is the dispatch. Regulators should look with skepticism on any proposal built on the flawed foundation that transmission usage and dispatch can be separated.

Restricting the Grid. The real reliability conditions for the electric grid include an ensemble of contingency conditions and complicated network interactions. Relatively few of these real constraints are simple limits on the actual flow across certain interfaces. Regulators should look skeptically at proposals that require derating the real capacity of the grid in order to make a few flow limits sufficient to guarantee reliability under a simple market model.

Pricing Transparency. Only the ISO would have the information needed to calculate and post locational prices, as in PJM. The computations are easy for a given dispatch, but only the ISO has all the information about the dispatch. Given the striking gap between the previous claims that congestion is insignificant and the observed reality of true locational pricing in the first real implementation in the United States, regulators everywhere should have a strong interest in prescribing that the real locational marginal costs--considering the real network interactions, and not just simplified zonal aggregations--be made available on a regular basis.

Regrettably, these illustrative recommendations are motivated by components found in quite real proposals, a few already in place. Some market participants may prefer large transaction costs, trading obscurity, barriers to entry, and the ability to exploit market power. They should oppose these and similar prescriptions which would simplify the real operations of markets and reduce the profits of those who otherwise would benefit from the inefficiencies. But we should not confuse the public interest in greater competition with an interest in greater profits for ever more competitors. Governments and regulators imbued with responsibility for the public interest should prevent such mistakes.

To be sure, embracing these recommendations would be the same thing as supporting an ISO that operates a "residual pool."[21] For some, the terms invoke a doctrinal reaction beyond reach of any reason. Despite all the evidence, for some the use of a system operator to coordinate an open spot market is politically incorrect. However, for those willing to look beyond preemptive dismissal and discuss the ideas on their merits, we should note that the equation runs both ways. Those who reject an ISO which operates such a voluntary, short-run, residual pool to coordinate the spot market are rejecting some or all of these principles, a rejection for which there seems to be little or no justification other than a fear of the open, competitive market that would result. Rejection of these principles amounts to saying that the ISO must perform certain functions, but badly.

Transmission Access and Pricing Challenges

Development of the rules for market design and access to the essential transmission facilities confronts a number of challenges. The basic outline of the bid-based, security-constrained economic dispatch with locational prices and transmission congestion contracts provides a foundation. If we could move beyond the distracting debates, and the avoid the avoidable design flaws, we could turn our attentions to an unfinished agenda. Some of the major, interconnected topics include:

> **Transmission Pricing for Fixed Charge Recovery**. Transmission displays large economies of scale, which means that the efficient way to recover embedded costs is in a system that separates fixed and variable charges. Recovery of fixed costs would further distinguish between those designed to recover sunk costs and those that would apply to new investments. There is no reason that sunk cost recovery should be uniform across the grid. This is a point of common confusion and is not the same thing as having a postage stamp rate. For the sunk costs, the license plate approach with different access charges for different regions--now in place, for example, in PJM, California and Australia--would be a preferred alternative, extending into the future. For new investment, fixed charges could be collected under a separate contract with those obtaining rights in the expanded system. Importantly, recovery of sunk costs has all the familiar characteristics of the stranded asset problem. In a world of truly open access, market participants would not willingly pay for the sunk costs. Hence, it is

[21] Richard Haigh, The National Grid Company, comments on Panel 5, "ISOs and Transmission Pricing," FERC Public Conference Concerning the Commission's Policy on Independent System Operators, April 16, 1998.

important (mandatory) that payments for embedded costs be mandatory.

Transmission Congestion Management System. The locational pricing system for managing transmission usage and congestion is central to the competitive market design. The structure of this pricing system supports the definition of transmission rights, incentives for investment, and the ability to rely on competitive market forces.

Transmission Rights for Energy. Under a market structure with bid-based economic dispatch and locational pricing, the natural definition of transmission rights is in the form of financial contracts for the collection of congestion rentals. The rights also serve as an important element of the incentives and obligations for the grid company. The rights are FTRs PJM or TCCs in New York. The basic design could be extended to include the explicit treatment of marginal losses and loss payments, as under consideration in New Zealand. Allocation of FTRs through the acquisition of network service creates certain perverse incentives in the re-designation process. An alternative is the development of auction mechanisms for allocating and reconfiguring transmission rights. Such auction proposals are under development in PJM and New York.

Transmission Requirements for Connection. There is a distinction between the requirements for new generators and loads to connect to the grid and the requirements for expansion of the grid. The connection rules will be important as both an obligation for the grid company and as an important source of future revenue.

Transmission Expansion Protocols. Separate from the connection to the grid are the incentives and rules for transmission system expansion in order to increase the capacity for energy trading and to create new FTRs. Here we would distinguish between market driven expansions and expansions dictated by market failure. For market driven expansions, the incentive for investment comes from the market participants who do not wish to pay future congestion costs and seek new FTRs. Or there might develop a mechanism that limited the working capacity of new investments for enough time to justify the return based only on ex post congestion rents. The market participants could approach the existing grid company and arrange for the investment, contract for future fixed payments, and receive the resulting FTRs. Presumably, any existing or new grid company could compete for this business, which could be contestable. For investments where market failure prevents the development of a transmission expansion, there would be a different but parallel decision mechanism. In this case, the grid company might take the initiative in identifying cost-effective expansion options that cannot be undertaken because of the free-rider problem blocking the formation of a sufficiently large coalition of beneficiaries. If the case can be established, the grid company could propose an allocation of costs and benefits. The costs would be collected in a manner similar to sunk costs for the existing grid. The benefits would be distributed in the form of incremental FTRs.

Transmission Rights for Capacity Credit. As long as there is an installed generation capacity requirement and market--as in PJM, New York and New England--there will be a need for a separate system of transmission capacity rights that would be similar in structure to the old notion of physical point-to-point rights. Although these rights would not be connected to the congestion payments, they would be defined according to a simultaneous feasibility condition for deliverability. The rights could be auctioned, tradable, and so on.

Transmission and Market Power. There is a potential interaction between transmission rights and the exercise of generation market power.[22] The basic point is that generators with market power could affect both the profitability of their generation and the value of any transmission rights they may hold.

Coordination Across Regions and Transmission Loading Relief. In a large interconnected grid, the issue of coordination across regions has important implications for the design and use of transmission rights, transmission expansion, and all the other aspects of the transmission business. There is a continuing over the procedures for transmission loading relief. A poorly designed TLR mechanism could undermine the market structure and severely reduce the value of the FTRs that can be viewed as the service provided by the wires company.

Ancillary Services. Market implementation problems have raised the level of concern about the treatment of the many activities that fall under the heading of ancillary services. Definitions vary, but many of these services could come directly from investments in generation or in the wires business, e.g. with capacitors to provide reactive power support. The policy focus on the energy and capacity markets should be balanced by some further investigation into the developments in ancillary services.[23]

Obligations of the Grid Company. The obligations to be imposed on grid companies have not been fully addressed in the context of the new competitive market designs. The England and Wales case is an exception, with its own problems. Elsewhere, the requirements for the Gridco have not been the focus of policy development. This is an area that could impose potentially large obligations on the Gridco, and where the complexity of the problems provides an opportunity to shift costs. A natural, market-oriented alternative would be to define the product in terms of the FTRs and other transmission rights created by the grid and then to define a set of financial obligations that would be connected, for example, to increased congestion costs.

---

[22] Judith B. Cardell, Carrie Cullen Hitt, William W. Hogan, "Market Power and Strategic Interaction in Electricity Networks," Resource and Energy Economics, 19(1997) 109-137.

[23] Eric Hirst and Brendan Kirby, "The Functions, Metrics, Costs, and Prices for Three Ancillary Services," Prepared for the Edison Electric Institute, October 1998.

Incentives for the Grid Company. The favorite subject of the grid owners is the financial incentives they should enjoy for maintaining the grid and expanding in a cost-effective manner. Such incentives would be appropriate if connected to the framework for obligations. For example, if investment in new grid facilities were contestable, and the obligations of the grid company principally were to stand behind the financial commitments in FTRs and other transmission rights, then a form of light-handed regulation would be possible, with cost based-rates for embedded cost recovery for sunk costs but negotiated market-based payments for some or all new investments. This is an area that has substantial potential, but the ideas are not yet well developed in the policy discussion.

This is not an exhaustive list, with its focus on market operations rather than governance and legal organization. But it does suggest the areas where progress is possible. The key to success would be to build on the fundamental economics of competitive electricity markets, and get the prices right.

GETTING THE PRICES RIGHT

The connection between prices and operating decisions often receives cursory treatment in the electricity restructuring process. Market participants want flexibility and choice, but object to consistent pricing as too complex. This is a mistake, and produces only an illusion of simplicity. If customers have flexibility in the choice of generation, spot purchases, bilateral transactions, and so on--then prices matter and competitive prices should reflect marginal costs. In large part, control of operating decisions is moving from engineers motivated by principles of technical efficiency, to market participants motivated by prices and profits. This is a major purpose of electricity restructuring--to change the locus of such key decisions. If we want the market to be guided by prices, and we expect and intend for people to take these prices seriously, it becomes important to follow the usual advice to "get the prices right." The experience in the first year with a consistent market pricing system in PJM underscores the point and provides a substantial database illuminating one of the central problems in electricity markets: pricing to allocate use of scarce transmission capacity.

In the United States, the move to a competitive electricity market with a consistent pricing system for allocating scarce transmission capacity entered a new phase beginning in April 1998 with the introduction of spot market locational pricing in PJM. The new system includes a spot market coordinated by the ISO. The ISO accepts both bilateral schedules and voluntary bids of the market participants. Using these schedules and bids, the ISO finds an economic, security-constrained dispatch for power flows and the associated locational marginal cost prices. Even without transmission constraints, this coordinated and transparent spot market provides significant benefits. When the transmission system is constrained, the spot prices can differ substantially across locations. Sales through the spot market are at the locational prices. The transmission usage charge for bilateral transactions is the difference in the locational prices between origin and destination. An accompanying system of FTRs provides financial hedges

between locations. These FTRs are the equivalent of perfectly tradeable firm transmission rights.[24]

In Markets with Choices, Prices Matter

The new PJM locational pricing system was embraced after an experiment during 1997 with an alternative zonal pricing approach that proved to be fundamentally inconsistent with a competitive market and user flexibility.[25] The experiment made the point in a dramatic way, as discussed further in the appendix. The important issue is not the total cost of transmission congestion, which may be small on average if the system is used efficiently, and when the cost is often mistakenly dismissed as irrelevant. Rather, the point is the incentives at the margin when the system is constrained. In designing the rules for transmission access and pricing for a competitive market, it matters little how the rules perform when the system is unconstrained. The important question is how the rules deal with the market and participant choices when the system is constrained. The earlier zonal pricing system allowed market participants the flexibility to choose between bilateral transactions and spot purchases, but did not simultaneously present them with the costs of their choices. The circumstances created a false and artificial impression that savings of $10 per MWh or more could be achieved simply by converting a spot transaction into a bilateral schedule. Faced with this perverse pricing incentive, market participants responded naturally by scheduling more bilateral transactions than the transmission system could accommodate. In effect, using the wrong prices induced behavior which greatly increased the cost of congestion. Inevitably, in June 1997 the ISO had to intervene by restricting the market and constraining choice to preserve reliability. The PJM ISO was fully aware of the perverse incentives of zonal congestion pricing and the problems they created. But without the authority to change the pricing rules, the ISO had no alternative but to restrict the market.

Even if the total cost of congestion might be modest over a year, a gap of $10 per MWh between the true costs of transmission usage and what participants pay is more than sufficient to get the attention of market participants at the time when it matters most, when the system is constrained. Given the margins in this business, market participants will change their behavior for $1. And the changes in behavior can substantially affect system operations; in fact, the whole point of electricity restructuring is that changes in behavior can affect system operations and lead to different patterns of electricity use and investment.

By contrast, the locational pricing system avoids this perverse incentive. By construction, the locational prices equal system marginal costs. Every generator would be

---

[24] An FTR is the same as a transmission congestion contract (TCC). For further details, see William W. Hogan, "Independent System Operator: Pricing and Flexibility in a Competitive Electricity Market," Center for Business and Government, Harvard University, February 1998, available on the author's web site.

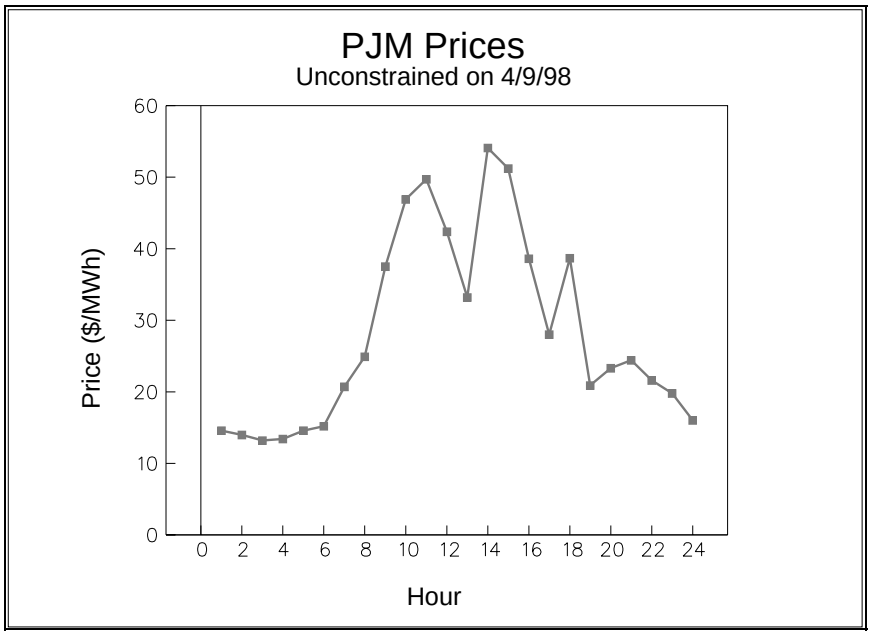
[25] Here the issue is pricing for transmission congestion. The recovery of embedded costs of transmission investment through access charges is a separate matter that is amenable to a zonal approach. Locational pricing has long been available in other markets, such as New Zealand. The PJM case is of interest because of its size, the sharp contrast of the debates, the experiment of trying two different pricing systems, and the availability of the data.

producing at its short-run profit maximizing output, given the prices. The market equilibrium would support the necessary dispatch in the presence of the transmission constraints. Spot-market transactions and bilateral schedules would be compatible. Flexibility would be allowed and reliability maintained consistent with the choices of the market participants.

Faced with this reality, the Federal Energy Regulatory Commission (FERC) acted to approve the locational pricing system that became operational in PJM at the beginning of April of 1998. The developing experience with this full locational pricing of the use of scarce transmission capacity deserves close study by the Commission and all system operators.

Transmission Congestion and Locational Prices

To put the problem in context, note that market clearing prices can vary substantially, even without transmission constraints.[26] The accompanying figure shows the prices in PJM over the day of April 9, 1998. This was not the most volatile day, and there were no transmission constraints during the day. However, the market clearing price varied from a low of $13 MWh to a high of $54. During June, the variation in unconstrained prices on the most expensive day increased by almost an order of magnitude to a high of $300, on June 26, 1998. As the summer continued, higher prices appeared even without internal transmission constraints in PJM. For example, on August 24 at 1400 and 1500 hours, the price reached the regulated maximum of $999. Clearly market participants must deal with substantial changes in prices, even without transmission congestion.[27]



Although these were the early days, the new locational pricing mechanism worked as anticipated by the ISO and the supporters of the approach, but apparently not as anticipated by

---

[26] The data used here were taken from the PJM web site at www.pjm.com.

[27] Prices outside of PJM reached even higher levels, reportedly as high as $7,000 per MWh in the Midwest; Wall Street Journal, June 29, 1998, p. C1. The apparent market disequilibrium between PJM and the Midwest is an important issue, but that is another story. The emphasis here is on the price implications for allocation of scarce transmission capacity within PJM.

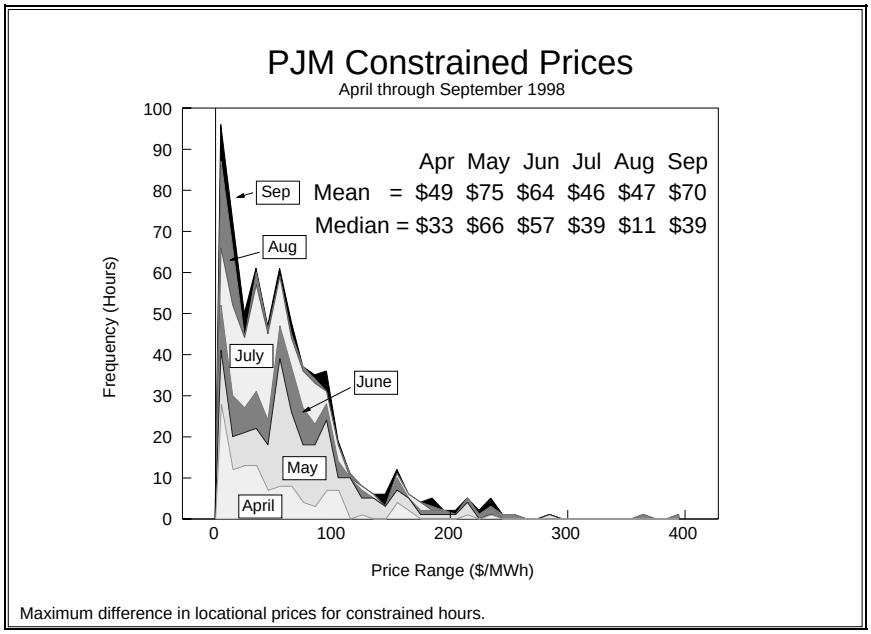
many who dismissed the importance of this issue. April and May are not typically highly constrained periods in PJM, and it would not have been impossible for the first days of locational pricing to have been boring. With no constraints, the locational prices, ignoring losses, would be identical at all locations. The cost of transmission between points--the difference in the locational prices--would have been zero. Nothing much might have happened until we approached the summer, when congestion would be more likely, as for the previous year in June.

In the event, the results were not boring. The system experienced transmission constraints, locational prices separated, and the opportunity cost of transmission was quite large. The lowest locational prices were sometimes negative, reflecting the value of counterflow in the system where it would be cheaper to pay participants to take power at some locations and so relieve transmission constraints. The highest locational prices were larger than the marginal cost of the most expensive plant running, reflecting the need to simultaneously increase output from expensive plants and decrease output from cheap plants, just to meet an increment of load at a constrained location. Over all hours in April 1998, for example, the low price was -$45 at 1500 hours on April 18 at "JACK PS," and the highest price was $232 at 1100 hours on April 16 at "SADDLEBR," both locations being in the Public Service Gas & Electric territory. Over the first year with the locational pricing system, the maximum difference between the lowest and highest contemporaneous prices was $412, at 1100 hours on November 19, reflecting the difference between $322 at "SADDLEBR" and -$91 at "BELLVIL." The second highest difference was $399, at 2000 hours on August 26, reflecting the difference between $437 at "ESAYRE" and $38 at "NYPP-W." This maximum price separation reached the same level as in the relatively unconstrained month of March 1998 before the locational prices were charged, when users could ignore the cost of congestion.[28]

The contemporaneous difference in locational prices, which is the price of transmission usage, has been large quite often. It does not take much of a difference to change behavior when the reported trading margins may be as low as $1 per MWh. If we take the $1 per MWh standard as an arbitrary threshold to define a constrained period, the range of highest to lowest price across locations exceeded the threshold for 119 hours in April, or approximately 17% of the time. As shown in the accompanying figure, the frequency distribution of the price range in constrained hours is skewed, with a median hourly price range at $33 and a mean of $49 in April. When the system is constrained and the market incentives matter the most, the marginal costs of transmission can be large indeed.

The monthly data for May through September, covering the summer peak, reinforce this initial impression. In general, May saw both higher prices and more transmission congestion. The difference between the highest and lowest locational price in May exceeded the $1 threshold for 183 hours, or approximately 25% of the time. As shown in the accompanying figure, the frequency distribution of this congestion price shifted to higher costs. In May, the median of the hourly price ranges doubled to $66 and the mean increased to $75. June was less

[28] On March 26, 1998, at 2200 hours, the difference between the highest to the lowest marginal cost was almost $400.



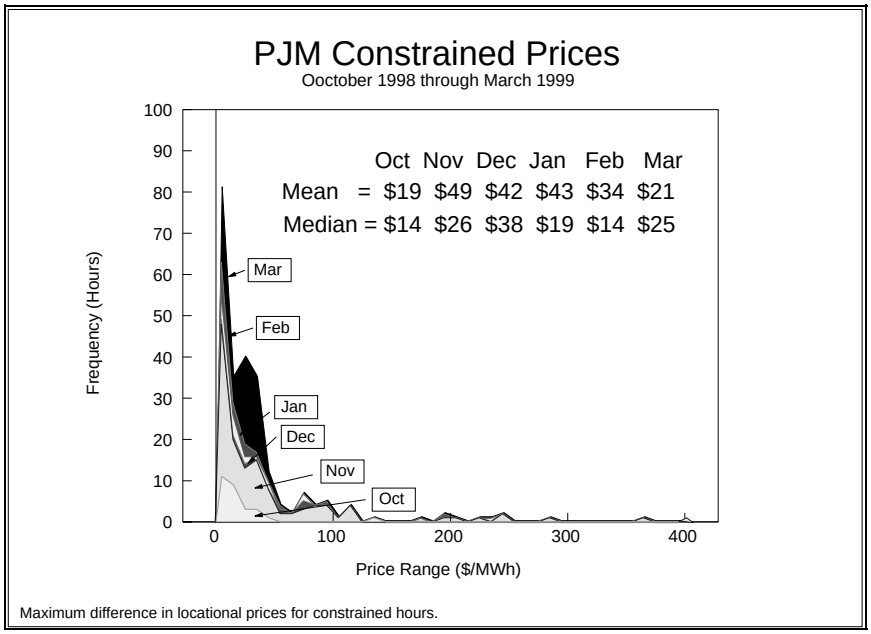
constrained, exceeding the $1 threshold for 95 hours or 13% of the time. The June median of the hourly price ranges was $57 and the mean was $64. During July, constraints appeared more often, as in May, with 151 constrained hours or about 20% of the time. July saw a median hourly price range of $39 and a mean of $46. By contrast, August showed locational constraints only for 48 hours or 7% of the time. The median hourly price range in August was $11 and the mean was $47, reflecting a few hours when the difference between the lowest to the highest price reached almost $400. September was like August, with 46 constrained hours or 6% of the time. However, the average price of congestion was high in September, with a mean of $70 and a median of $39.

The experience of higher unconstrained prices and fewer constrained hours in June, August and September reminds us that the period of peak system load is not necessarily the time of greatest transmission congestion. Transmission congestion reflects an imbalance in the location of load and generation. At peak load, more generation comes on line and may relieve system congestion. In addition, the particular flow of power into the Midwest, reversing the usual direction, tended to unload the transmission constraints during the summer of 1998.

This record of continuing constraints was reinforced by the events in the following months from October 1998 to March 1999. After the heavier loading of the PJM summer, the winter months would be less constrained but the constraints did not disappear. As shown in the accompanying figure, the frequency diagram of price ranges showed that some significant constraints applied. November alone accounted for 105 of the 242 constrained hours over the period. The median price range for the constrained hours in November was $26 and the mean was $49. The corresponding median and mean price ranges of the other months for the hours that the system was constrained appear in the figure.[29]

The evidence shows many things. For example, calculating and reporting the locational prices for each point on the grid are not especially complex tasks, at least for the system operator

[29] The numbers of constrained hours for October 1998 through March 1999 were 28, 105, 6, 20, 18, and 65, respectively.



who has the necessary information available. The prices can be available every five minutes on the Internet. Faced with these prices, the market participants adjust their behavior, just as intended. The transition was not painless, especially for those who ignored many warnings and entered into "seller's choice" contracts that gave the seller the maximum theoretical financial advantage for relieving congestion. Presumably, this form of contract will disappear, or be properly priced in the future, and market participants will become more attuned to the use of fixed transmission rights to hedge much of the cost of congestion. But market participants who rely on the spot market, and are not prepared to pay for congestion hedges that fix the cost of transmission in advance, will see price signals that align their incentives with the reality of system operations.

Full locational pricing is fully compatible with a trading system built on a hub-and-spoke framework. The hub becomes a common trading point, and the cost of moving to and from the hub, along the spokes, is just the difference in the locational prices.[30] If the nodal prices are available from the ISO, market participants can define their own hubs. In the PJM case, however, market participants asked the ISO to handle the accounting to create several hubs, of which the western hub has so far developed as the preferred trading point.

The full market response to all these changes is not known because the data are not all in the public domain. However, one information source is a sampling of trader activity reported in the Wall Street Journal.[31] According to these data, the immediate response of the market was to reduce reported spot trading in April of 1998. However, by mid-May of 1998 reported transactions had returned to volumes comparable to those seen just before the new locational pricing system went into effect. Subsequently, and reversing its earlier objection that the nodal pricing market would not be sufficiently liquid, in March of 1999 the New York Mercantile Exchange launched a new futures contract to capitalize on the highly liquid trading market that had developed at the PJM western hub. Apparently inadequate liquidity was not a

---

[30] See the appendix for a further discussion.

[31] For example, see "DJ Electricity Price Indexes," Wall Street Journal, June 3, 1998, p. C19.

problem. Even further, the spot and forward markets at the western hub were reported to be so liquid that the futures contract might not be able to compete.[32] Although market liquidity is often vaguely defined and seen only in the eye of the beholder, as reported by these sources the market appears to have adjusted to the new environment within a framework that supports transactions with consistent prices.

The operational problems experienced by the ISO in the year before full locational pricing, where profit driven market participants undermined reliability, did not appear in the year after adoption of full locational pricing. Locational pricing presents profit driven market participants with the right incentives consistent with the true opportunity costs. This same pricing system was applied by PJM for managing inter-regional transmission loading relief. With full locational pricing, the prices reinforce reliability. In addition, the anecdotal evidence suggests that investments in new generation and transmission were being considered with careful attention to the effects of system congestion, just as intended.

In the first year, generators within PJM's boundaries faced bidding constraints intended to impose competitive behavior. Generators outside of PJM selling into the market had no bidding restrictions, and often set the market-clearing locational prices. Based on a conclusion that there is no significant market power within PJM, at the start of the second year after application of full locational pricing, generators inside PJM also received FERC authorization to go to full market-based bidding. The experience with the first year and the success of PJM locational pricing will provide one useful benchmark for comparing the performance of the market without bidding constraints. As discussed in the appendix, if there is market power in PJM, full locational pricing should help mitigate that power.

Full Locational Pricing is the Truly Simple Approach

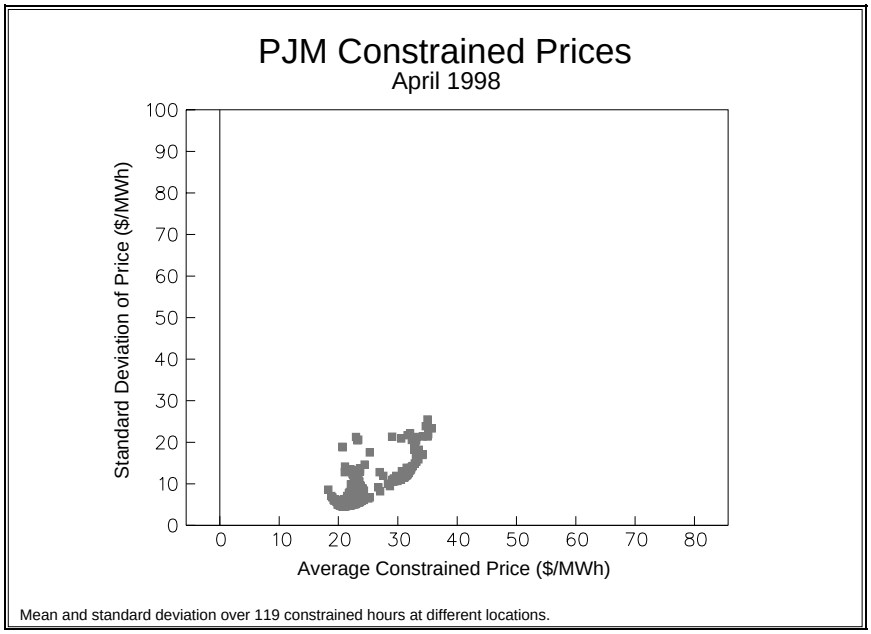
What about aggregating PJM into a few zones, if not just one? The PJM ISO is providing prices for approximately 2000 locations. This is a convenient way to represent the information, because it is how the data are organized for actual system operations. However, some of these locations are really just multiple units at the same point on the grid, and would necessarily have the same prices in most circumstances. For other points on the grid, the zonal argument is that the locational differences would be minor, and could be represented by a relatively few zones. This view has been subjected to a test over the first year of operation.

The period April through at least early June could have been relatively unconstrained, presenting a low hurdle for the zonal approach. The choice of the appropriate zones would not be an easy matter, and there is some ambiguity in clustering criteria. However, one simple way to summarize the data would be to examine both the average and the variation of prices at different locations during constrained hours. If two locations always have the same prices, then

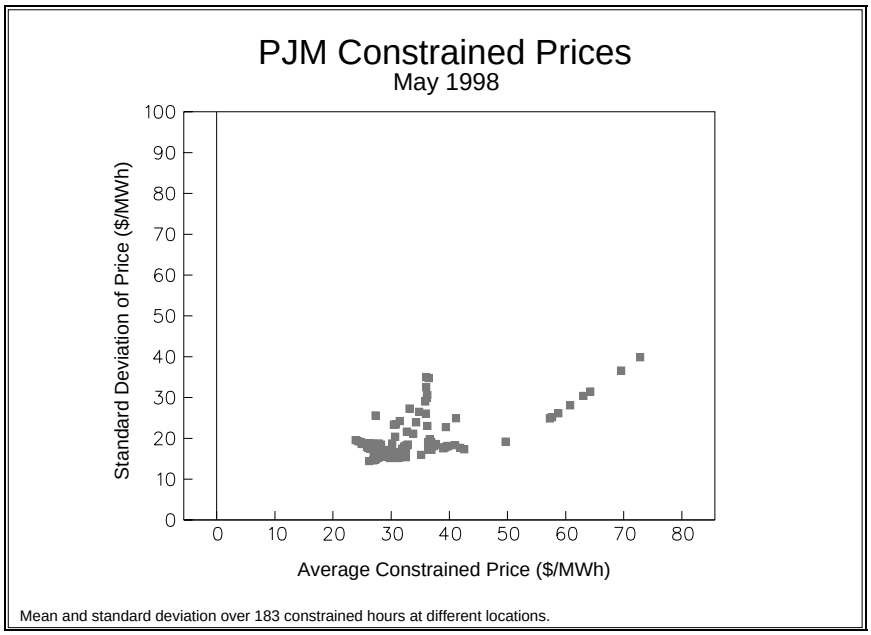
---

[32] "The New York Mercantile Exchange will launch an electricity futures contract March 19 at the PJM western hub, one of the most liquid markets in the Eastern grid. ... The PJM hub already features an active and growing over-the-counter forwards market. A liquid hub can have a downside [for the futures contract] given that players are content trading in the OTC, said one Northeast broker." Power Markets Week, February 8, 1999, p. 14.

the two averages of prices over the period would be the same and the two standard deviations of the prices would be the same. These conditions would be necessary, but not sufficient, for the prices to be the same at the two locations. Hence, this straightforward calculation gives a lower bound on the number of different locations with sometimes unique prices.



Mean and standard deviation over 119 constrained hours at different locations.

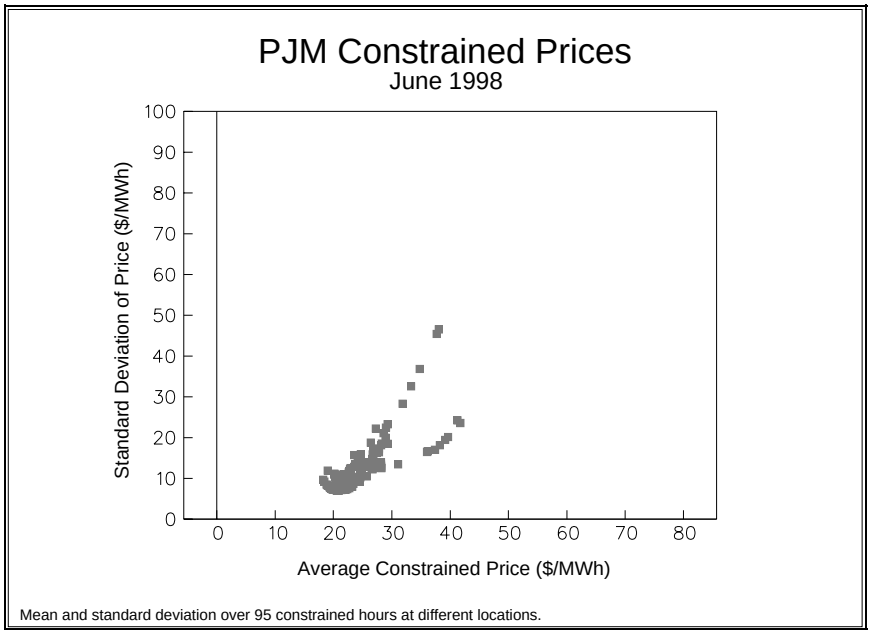


Mean and standard deviation over 183 constrained hours at different locations.

The accompanying figure for April plots the data on average price and standard deviation of price across 119 constrained hours in April for all the locations reported by the PJM ISO. There are 2000 points in the graph, one for each location. Were it true that there were only a few zones, the graph would show a few clusters of locations where the average prices were the same and the standard deviations were the same. In fact, there is substantial dispersion.[33] After the first month of operation, there were 766 locations within PJM where the price points did not overlap and were different by this lower bound test. The corresponding data for May show a similar dispersion and higher costs of congestion. The accompanying figure plots the locational means and standard deviations for May, for which there were 789 different locations by this lower bound test. There were not as many limiting constraints, but even one constraint can produce

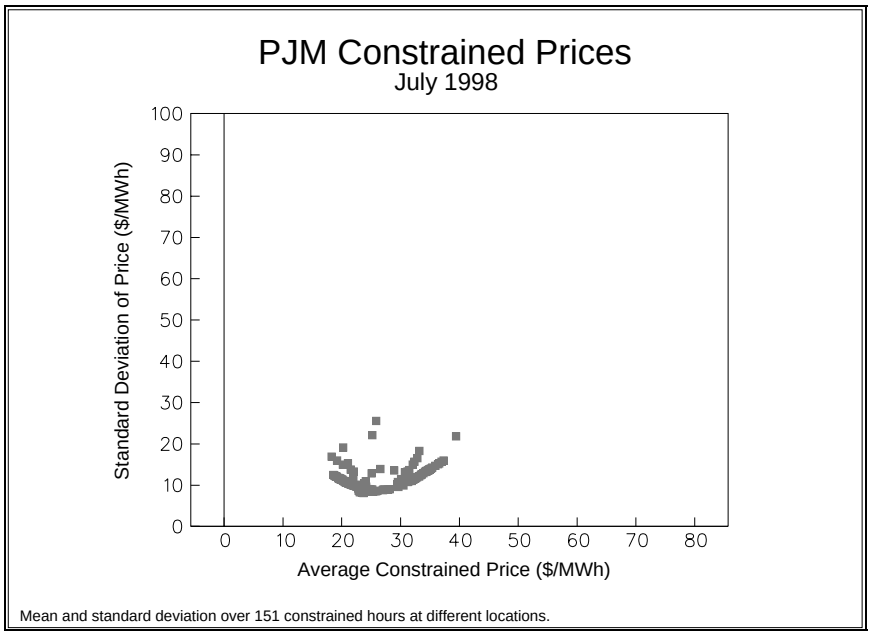
[33] The large scale in the graphic allows for consistent comparison with later months, when the price dispersion is even greater.

a substantial range of prices.



Mean and standard deviation over 95 constrained hours at different locations.

In June there was a similar dispersion of prices. As shown in the corresponding figure for June, there were 689 locations with unique prices according to the lower bound test of having different averages or different standard deviations. The data for July in the subsequent figure show more congested hours and more locations, with a total of 785 different points, much as in May. By contrast, August revealed relatively fewer congested hours. Nonetheless, even in August the data indicate 693 different locations. September showed greater price dispersion with 825 different locations appearing in the price data.



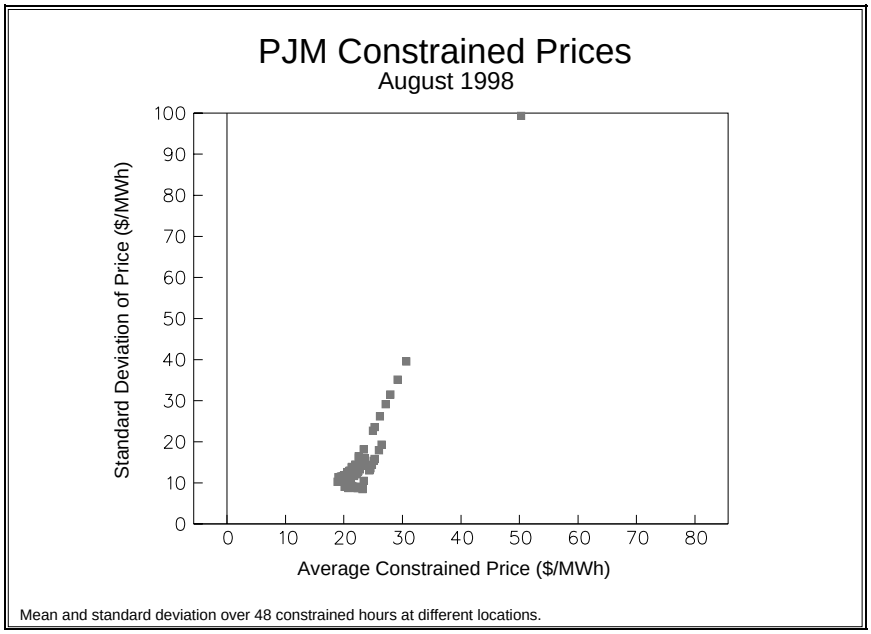
Mean and standard deviation over 151 constrained hours at different locations.

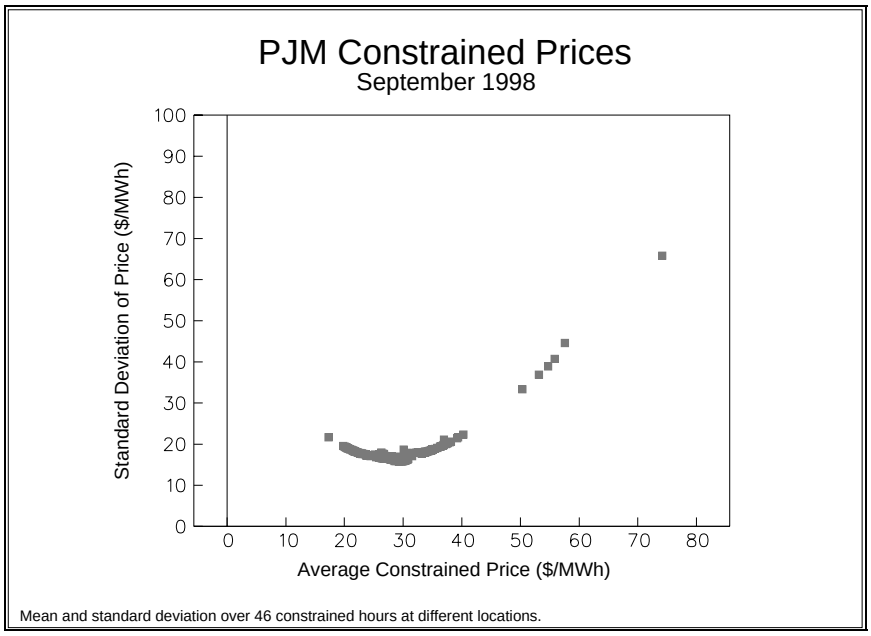
Furthermore, the differences were not the same across the months. The publicly available data for June through September cannot be merged easily with the previous months due to changes in the names and number of locations. However, using pooled data that calculates the averages and standard deviations for each location across both April and May, there were 821 different locations where either the average price or the standard deviation differed from other locations. Presumably as more experience develops and different constraints appear, this number will grow. Recall that in a sufficiently interconnected network, a single thermal limit on a transmission line could create different prices at every location, sometimes very different prices. This is contrary to the intuition that arises from the misleading analogy to a simple radial transmission connection without any network interactions, where a single constraint results in

only two prices. But the network interactions and the many different prices are quite real and no surprise to the system operators.

The criterion of no difference in prices may be too strict, and we might be willing to declare two locations as the same if they are close enough. The criterion should be set so that the maximum difference in prices across a zone would be small enough not to affect behavior. Note that this is not the same having a small contemporaneous standard deviation of the prices across the zone, which would leave many locations with the perverse incentives that create the complicated rules for market interventions.

For instance, consider the criterion proposed to select a zone such that the standard deviation of prices across the zone is less than 10% of the average prices.[34]





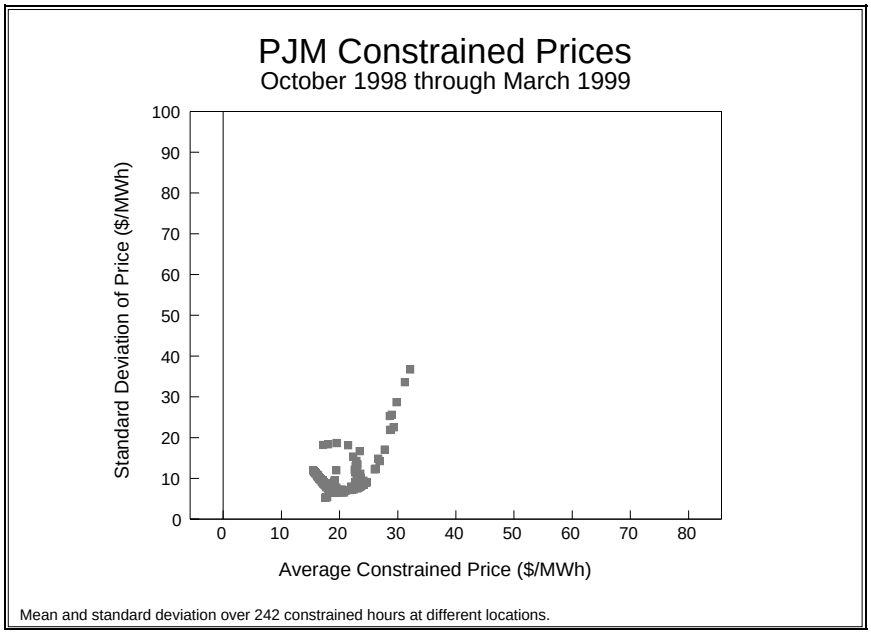
[34] This 10% standard deviation rule across a zone was adopted in Richard D. Tabors, "Transmission Pricing in PJM: Allowing the Economics of the Market to Work," Tabors Caramanis & Associates, February 24, 1999, p.17. Even with this relaxed rule, there are major regions in PJM where the zonal approach would fail, notably in northern New Jersey during the period April 1998 to September 1998. The analysis illustrates some of the problems with the zonal approach. For example, consider the constraint "shift factor" evaluation that suggests only a few zones for PJM. The data show that the 28 constraints examined (Tabors, p. 31.) include *none* of the 43 constraints that actually limited transmission during the first six months of locational pricing, as reported on the PJM web page.

Apparently the view is that if this 10% criterion were met, *and market participants did not respond to the incentives*, cost shifting would be small. However, with average prices of $25 and normal distributions, this would mean that at least one third of the locations within a zone would see price differences of more than $5 per MWh, surely enough to cause the same problems that PJM experienced. The problem is that the participants *will* respond to the incentives and this will create the inevitable need for the ISO to adopt arbitrary rules to restrict the market. The goal should be avoid the perverse incentives, and not have to hope that market participants will ignore trading opportunities.

Defining the standard for "close enough" could be contentious, but it may be moot. If we accept the $1 per MWh threshold above for the maximum deviation of prices across a zone, and ask how many separate zones would be necessary to cover all the points in the figures, the answer is 94 zones in April, 83 zones in May, 75 zones in June, 57 zones in July, 52 zones in August, and 64 zones in September.[35]

Again, these are not the same zones in each month. If we pool the months of April and May and apply the consistent threshold of $0.50 per MWh average difference over two months, we find 132 different zones needed to capture the variability in locational prices, so far. This is many more than the few zones predicted, and there is no reason to believe that we are finished adding to the list.

For the months of October through March, the data are easier to combine across locations. The corresponding presentation of the average prices and their variation appears in the accompanying figure. For these six months, there are 842 different locations with distinct prices. Applying the same $1 threshold, these would aggregate into 61 zones. The number of zones that would be required in any given month, which range from a low of



[35] Starting with separate zones for the ten service territories and sorting on the average and then the standard deviation. If we do not require the individual service territories to be separate zones, the number of distinct price zones is 60 in April. The corresponding figures for a $5 per MWh threshold would be 28 and 10 zones in April, with and without separating service territories.

38 in October to 71 in January.[36] Note that the number of zones required is not solely a function of the number of constrained hours. For example, in December, with only 6 constrained hours, the constraints were such that 64 zones would have been required to capture the variability in price. In short, a few zones wouldn't do, many would be required.

Furthermore, as is obvious, the effect of locational pricing for "natural" zones would be to produce the same price for all such locations. Hence, aggregating to zones where prices are almost the same would result in many zones and no apparent simplification with any meaning. We might as well do the simpler thing of using the locational prices at each location. Aggregating across the many real locations to a few zones, by contrast, necessarily means combining like with unlike, thereby recreating in microcosm the perverse incentives of the failed experiment with a single zone, leading to a breakdown of the non-discriminatory market and administrative restrictions on choice.

The interaction between reliability (with its inescapable physical realities) and economics will limit the acceptable ISO access and pricing rules for allocating scarce transmission capacity. It would be desirable to offer market participants flexibility in their own decisions. A great deal of flexibility in combining a range of bilateral schedules and spot transactions would be possible. However, the more varied and flexible the options for the market participants, the more important it will be to get the prices right, meaning consistent with the marginal impacts on the system. The whole point of the turn to greater reliance on competition is that the market participants will respond to incentives. As we have seen, if prices don't provide the right incentives, consistent with the impacts on the system, the participants will respond in their own interests without concern for the system effects, and the ISO will be driven inexorably to intervene in the market and restrict choice. The locational pricing system with FTRs works in PJM, and a similar system received FERC approval for the New York ISO.[37] With locational pricing, participant incentives are aligned. Buyers and sellers can buy and sell as they choose through the spot market at the locational prices. Or they can schedule bilateral transactions and pay the difference in locational prices as the charge for transmission usage. The result is flexible, non-discriminatory, and compatible with the mandates of reliability.

GETTING THE PRICES WRONG

The experience of PJM from 1997 and 1998--first getting the prices wrong, and then turning to locational pricing to get the prices right--is instructive but not unique. The PJM experience has the advantage of a sharp contrast and an easily available set of data that support analysis and illustration. However, other markets have pursued "simple" market design solutions

---

[36] The numbers of zones for October 1998 through March 1999 were 38, 67, 64, 71, 62, and 65, respectively.

[37] Federal Energy Regulatory Commission, New York ISO Ruling, Docket Nos. ER97-1523-000, OA97-470-000 and ER97-4234-000, January 27, 1999.

that avoid the basic model with bid-based, security-constrained economic dispatch and nodal pricing. These simple solutions soon create problems. Market participants respond to the misplaced incentives. In further response, the market designers or system operators must create rules to constrain market behavior. The resulting systems are more complex and lead to more regulation with less reliance on the market.

Often the connections among market design, market incentives, and market behavior are murky or missed. Introduction of ad hoc rules to fix one problem, creates others. A return to fundamentals becomes more difficult because a change in the rules creates winners and losers. Hence, the flaws of the basic design persist, often without even recognition of their role in complicating the market. Here a summary of illustrative examples from other markets illustrates the point.

New England and Barriers to Entry

New England operated a power pool (NEPOOL) for many years with an intricate but hidden pricing system that shared the costs among all the participants. The basic economic dispatch did not allow for choices in operations, and a complicated planning process required transmission investment along with generation to minimize the impact of new generators on the older units. From the perspectives of the NEPOOL companies, the network system appeared to be largely unconstrained within New England, and this seemed to be an advantage in keeping things simple.

In setting up the design to reform the pool for a restructured NEPOOL market, the designers pursued this goal of simplicity. Although the details are interesting, the core idea was simple. In effect, NEPOOL adopted the same one-zone pricing system that PJM pursued in 1997.[38] The simple zonal system assumed no transmission congestion, or that any congestion would have a minimal effect. In practice, this may never have worked for the existing NEPOOL network, once the forces of the market had allowed the incentives for changing the pattern of use of the available generators. However, the assumption would not be true with the introduction of new generators, at least not without substantial expansion of the transmission grid.[39]

In response to the market opportunities in New England, approximately 30,000 MW of new plant construction was announced. If built, this would more than double the capacity of the system. However, under the simple pricing system, the new generators would have every incentive to build while ignoring the costs in transmission congestion. Recognizing that this would be untenable, NEPOOL proposed a complex system of rules for new generators, requiring both extensive studies of system impacts and expensive investments in the transmission system. One effect would have been to impose substantial costs and long delays on the entry of new

---

[38] In the complicated process of setting up a market design that satisfies the many stakeholders, this is evidence of how hard it can be to learn from the experience of others.

[39] For a critique of the NEPOOL one-zone congestion pricing system, see Peter Cramton and Robert Wilson, “A Review of ISO New England's Proposed Market Rules,” Market Design, Inc., September 9, 1998.

generation. Another would have been the protection of existing generation from the full force of competition.

In October 1998, the Federal regulators struck down these barriers to entry.[40] In effect, the FERC said that the neither the regulators nor the new generators bore responsibility for the incentives created by the defective pricing system proposed as a tool for congestion management. The rules for limiting entry of new generators were in effect both discriminatory and anti-competitive. Something else would be required. While not going so far as to mandate a particular congestion management system, FERC's actions made it impossible to maintain the old "simple" solution, which of course turned out not to be so simple after all.

Immediately, NEPOOL began struggling with developing an alternative congestion management system. One option is to put a locational pricing system in place. The previous one-zone system captured the worst features of the failed experience in PJM. The simple solution provided the wrong incentives, and the market responded. The response threatened to overwhelm the network. The lesson from PJM in favor of the simplicity of a true nodal pricing approach could now be applied in New England, and such a proposal was submitted to the FERC on March 31, 1999.

Next New England will have to look to the design problems created by importing a design flaw from California in the separation of its markets for ancillary services and energy.

California and Loss of Economic Dispatch

California embraced the design principle that whenever possible, and even when not, markets should be separated and segmented. This included the separation of the spot market operated through a PX and the activities of the ISO. Furthermore, the several categories for ancillary services, such as spinning and non-spinning reserve, replacement power, and so on, were set up with their own separate markets. Transmission congestion pricing was applied to a number of zones, but the rules are ad hoc and opaque.

Unlike PJM, California did not have the existing infrastructure of a power pool to use in launching the new market design. Hence, the first year of operation, commencing March 31 in 1998, saw a number of start up problems. At the instigation of the FERC, the first year also saw an extensive set of analyses of the developing problems. Although a full review of the experience in California is beyond the scope here, at least two points bear on the fundamental design of the market institutions.

First, markets are segmented and treated separately, but the underlying technology means that the same machines can and must provide the separate services. Hence, a generator can use its capacity to supply energy, or spinning reserve, or replacement reserve. But the total cannot exceed the actual capacity. In markets for most commodities, this jointness in production

---

[40] Federal Energy Regulatory Commission, New England Power Pool Ruling, Docket No. ER98-3853-000, October 29, 1998.

would present no fundamental problem. The several products could be offered in separate markets, with temporary imbalances buffered by inventories or eliminated by changing prices. An oil refinery can produce gasoline and heating oil, without requiring a coordinated market for these products.

Unfortunately, this is not the case for electricity. There is no inventory to provide a buffer, and there is no time to adjust supply and demand through the price mechanism. Despite the advance claims of the proponents of decentralized markets as the solution to all problems, not just some problems, the markets could not and did not work as anticipated. The ISO was precluded from recognizing and accommodating the inherent interactions in production. Generators were required to guess in advance as to how much spin and how much replacement reserve would be offered by others. If they guessed wrong, they could lose in the energy market and sell their capacity for a $5/MW in the replacement reserve market. Or they could, and some did, win the lottery and sell that capacity in the same replacement reserve market for $10,000/MW. This surprising outcome may not have been quite as dramatic as the possibility of the lights going out in PJM in June of 1997, but it was dramatic enough. Reform is now in process, with tortured language to hide the fact that the answer lies in using a bid-based dispatch coordinated through the ISO.[41]

The problems of market separation and simple pricing in California have not been limited to the market for ancillary services. In order to preserve the separation of the ISO and the PX, it was necessary to keep the ISO from doing what comes naturally. Hence the rules specified restrictions such as the requirement that scheduling coordinators maintain balanced schedules, unlike everywhere else in the world where the only requirement is to maintain aggregate balance. This requirement to balance individually can significantly complicate the dispatch problem for the ISO, who in the end must ensure balance of the system as a whole. Furthermore, since the rule is extended to preclude trading among participants from being coordinated through the ISO, it has the effect of using the ISO to enforce the wishes of market participants who want to exercise market power and restrict competition.[42]

Other restrictions further preclude the ISO from implementing a least-cost dispatch. Bids are modified or rejected, precisely to prevent market participants from trading through the ISO market without going through the PX. However, this has not prevented market participants from following the natural incentives, and producing what appears to be anomalous or pernicious behavior. Furthermore, there is an increasing reliance on reserves and reliability-must-run (RMR) contracts to deal with the limitations imposed on the ISO, who in the end must meet the constraints in the real system.

---

[41] For example, see Frank Wolak, Robert Nordhaus, and Carl Shapiro, "Report on the Redesign of Markets for Ancillary Services and Real-Time Energy," Market Surveillance Committee of the California Independent System Operator, March 25, 1999.

[42] Paul R. Gribik, George Angelidis, and Ross R. Kovacs, "Transmission Access and Pricing with Multiple Separate Energy Forward Markets," IEEE Conference, Tampa, February 1998, pp. 1-2.

The pricing system for congestion within zones in California creates the same perverse incentives that we saw in New England. Generators would maximize their profits by locating plants at sites that would increase system congestion, because the generators do not see the direct costs of congestion which are socialized in the zonal response. In response, as in New England, the California ISO was reported to be considering entry restrictions on generators. Presumably this will receive a similar reception at FERC, rejecting the restrictions and reinforcing the need for a better congestion management system; there is a simple answer with integrated bid-based economic dispatch and locational pricing.

Australian Pursuit of "Firm" Rights Between and Within Regions

The Australian market is in the process of operating under a new ISO that handles the dispatch throughout the country. The National Electricity Market Management Company operates a bid-based economic dispatch. There are no formal provisions for separate scheduling of bilateral transactions, but these could operate in effect through the bidding system.

The existing NEMMCO pricing system is a mixture of zonal and nodal differentiation. Loss prices differ by node, but congestion is treated on a zonal basis, with large zones defined by states such as Victoria and New South Wales. The pricing system was created in the interest of simplicity, and in the mistaken belief that it would help in mitigating market power.[43] However, as part of an extensive review of the transmission pricing arrangement, the basic market design on a zonal system is a subject of discussion and possible reform.[44]

One difficulty that has appeared in the Australian market is in the lack of a workable system for defining transmission rights. For the same fundamental reasons that make zonal prices a very imperfect approximation of the true opportunity costs in a constrained transmission system, namely because of the effect of loop flow, it is impossible to define the transmission capacity of the system between broadly aggregated zones with looped interconnections. In effect, the lack of full nodal congestion pricing makes it impossible to define inter-regional hedges or transmission rights for the full capacity of the inter-regional connections. The true capacity depends on the pattern of usage, reflecting the same facts that give rise to different nodal congestion costs.

The problems are compounded within regions, where the cost of congestion that differentiates locations is socialized. Within these zones, generators cannot be guaranteed to run and receive the regional price, even though it would appear profitable. Generators who do not run are not paid, and they do not have firm rights which would be naturally available in a nodal pricing system. It is not possible to define the firm rights, because loop flow dictates that the pattern of feasible rights would depend on the pattern of use of the system. With true nodal

---

[43] See the appendix for further discussion on the point of zones enhancing rather than mitigating market power.

[44] National Electricity Code Administrator (NECA), "Transmission and Distribution Pricing Review," Draft report, Australia, March 1999.

pricing, this problem would be handled through the financial FTRs, as in PJM. But without nodal pricing, the seemingly simpler market confronts the inability to define intra- and inter-regional transmission rights that use the full capacity of the system. Further, it is difficult to define the benefits that would accrue from transmission expansion.

There is a simple reform, which would be a modest change in Australia. The bid-based, security-constrained economic dispatch of NEMMCO already includes locational losses and congestion between zones and losses within zones. All that is needed is to extend the system to include congestion pricing at all locations, with the introduction of FTRs to provide the basis for improved transmission rights and incentives.

England and Wales and the Pool Reforms

The major innovations of the market in England and Wales did not include a system for dealing efficiently with transmission congestion. In effect, the model is a single zone system for usage, with a partial socialization of the costs of transmission constraints. The system is now subject to extensive review and possible reform, although the best analysis suggests that the reforms will fix the part that is not broken.[45]

Importantly, the England and Wales pool has a zonal pricing system with a fundamental difference from that in PJM. Before nodal pricing applied the correct incentives, under the original zonal pricing system in PJM did not compensate generators who were constrained off. This zonal pricing approach created the strong incentive to run anyway, and the need for rules that prevented the generators from following their profit incentives. By contrast, the constrained-off generators in England and Wales are paid their opportunity costs, defined as the profit they would have made if the system had been truly unconstrained. This removes the incentive to deviate from the economic dispatch in the same way that U.S. farmers have long been compensated to follow the production plan; the generators are paid not to run, just as the farmers have been paid not to grow crops.

The farm subsidy analogy suggests the perverse long-run incentives that must follow. In England and Wales, there are no short-term congestion costs for generators, who make money whether or not they generate power. The only congestion signal by location, therefore, has been in connection costs. But for both analytical and political reasons, it is difficult to estimate the efficient connection cost before the fact, and even more difficult to pass on the full cost in the form of connection charges with no tradeable transmission rights. As a result, connection charges have been inadequate and there has been the predictable behavior of subsidized new generators building too much and siting where it is uneconomic:

> "The clustering of gas-fired power stations around the East coast of England, close to the onshore gas terminals in order to minimize gas infrastructure costs, is likely to cause electricity Grid System infrastructure constraints in

[45] Richard Green, "Draining the Pool: The Reform of Electricity Trading in England and Wales," University of Cambridge, December 1998.

> East Anglia. Similarly, infrastructure constraints could arise with gas-fired generation in the North West of England. Under the existing charging arrangements the costs of Grid System reinforcements to overcome constraints are passed on to all electricity consumers. The incentives to optimize the location of generation plant to minimize Grid System reinforcements are not strong enough to overcome the advantage of siting generating plant close to the source of fuel rather than closer to the demand."[46]

The market simplification, therefore, reinforced the dash to gas. The subsidies contributed to extensive introduction of new gas plants. These in turn idled otherwise economic coal plants with their politically important coal workers. Soon the government intervened, and imposed a moratorium on the construction of new natural gas plants.[47] Although many forces reinforced the moratorium, not just pricing, the incentives were important. Once again, bad market design and bad pricing turned out not to be so simple.

CONCLUSION

Public policy development in the continuing evolution of electricity restructuring emphasizes the institutions for market operations. Interconnections through the transmission grid create the necessity for regional organizations that can accommodate competition in services, generation, and contracting while preserving the reliability of the transmission system. Alternative models are many, but the developing experiences around the world provide insight into the options and implications of alternative models. Comparison of these models provides further information about details of market operations. It is apparent from this experience that there must be a close connection between the design of options for market flexibility and the pricing principles for use of the transmission grid. If market design and prices closely conform to operating conditions and marginal costs, then market participants can have numerous choices in the way they use the transmission system. However, if pricing does not conform to the operating conditions, then substantial operating restrictions must be imposed to preserve system reliability. Customer flexibility and choice require efficient pricing; inefficient pricing necessarily requires limits on market flexibility.

Examples of failure and success illustrate this close connection between pricing provisions and operating rules. Case studies from California, Australia, England and Wales, the New England Power Pool, and the Pennsylvania-New Jersey-Maryland Interconnection illuminate

---

[46] Merz and McLellan and Advanced Engineering Solutions, "Review of Energy Sources for Power Generation: Electricity System Study: Final Report for the Department of Trade and Industry," 16 June 1998, p. 8.

[47] Department of Trade and Industry, "Conclusions of the Review Process of Energy Sources for Power Generation and Government Response to Fourth and Fifth Reports of Trade and Industry," London, October 1998, p. 12.

the basic point. The extensive experience of the first year of full locational pricing in PJM illustrates the importance of the real locational marginal costs. The network effects can be surprising for virtually everyone other than experienced system operators. The conclusion points to the integrated independent system operator, de jure as an independent entity or de facto as a separate group within a larger Transco, with bid-based, security-constrained economic dispatch and locational marginal cost pricing rules, as the model most likely to be successful in preserving system reliability while supporting competitive markets with customer choice.

APPENDIX

The Nodal-Zonal Debate

Full locational pricing at every node in the network is a natural consequence of the basic economics of a competitive electricity market. However, it has been common around the world to assert, usually without apparent need for much further justification, that nodal pricing would be too complicated and aggregation into zones with socialization of the attendant costs would be simpler and solve all manner of problems. On first impression, the argument appears correct. On closer examination, however, we find the opposite to be true, once we consider the incentives created by aggregation combined with the flexibility allowed by market choices. But the debate continues.

For example, the original one-zone congestion pricing system proposed for the New England independent system operator (ISO) created inefficient incentives for locating new generation.[48] To counter these price incentives, the proposal imposed limiting conditions on new generation construction. Following the FERC rejection of the resulting barriers to entry for new generation in New England, there developed a debate over the preferred model for managing and pricing transmission congestion.[49] One zone was not enough, but perhaps a few would do? Or should New England go all the way to a nodal pricing system as in PJM?

> **Fact:** A single transmission constraint in an electric network can produce different prices at every node. Simply put, the different nodal prices arise because every location has a different effect on the constraint. This feature of electric networks is caused by the physics of parallel flows. Unfortunately, if you are not an electrical engineer, you probably have very bad intuition about the implications of this fact. You are not alone.
>
> **Fiction:** We could avoid the complications of dealing directly with nodal pricing by aggregating nodes with similar prices into a few zones. The result would provide a foundation for a simpler competitive market structure.

There are many flaws in this seductive simplification argument.[50] In reality, the truly

---

[48] The use of zones for collecting transmission fixed charges is not the issue here. The focus is on managing transmission congestion. For a critique of the previously proposed one-zone congestion pricing system, see Peter Cramton and Robert Wilson, "A Review of ISO New England's Proposed Market Rules," Market Design, Inc., September 9, 1998.

[49] Federal Energy Regulatory Commission, New England Power Pool Ruling, Docket No. ER98-3853-000, October 29, 1998.

[50] William W. Hogan, "Nodes and Zones in Electricity Markets: Seeking Simplified Congestion Pricing," in Hung-po Chao and Hilliard G. Huntington (eds.), Defining Competitive Electricity Markets, Kluwer Academic Publishers, 1998, pp. 33-62. Steve Stoft, "Transmission Pricing in Zones: Simple or Complex?", The Electricity

simple system turns out to be a market that uses nodal pricing in conjunction with a bid-based, security-constrained, economic dispatch administered by an independent system operator. Purchases and sales in the balancing spot market would be at the nodal prices. Bilateral transactions would be charged for transmission congestion at the difference in the nodal prices at source and destination. Transmission congestion contracts would provide price certainty for those who pay in advance for these financial “firm” transmission rights up to the capacity of the grid. The system would be efficient and internally consistent.

The debate over transmission usage and the earlier experiment with a zonal transmission pricing approach in PJM provide a stark illustration of the difficulty and the challenge.[51] In March of 1997, the FERC approved an interim transmission access and pricing system to operate in conjunction with a real-time spot market coordinated through the PJM ISO. Faced with opposition to a full locational pricing and congestion charging mechanism for actual use of the system, the FERC endorsed the locational approach in principle but adopted temporarily an alternative model proposed by Philadelphia Electric Company (PECO) and others. The PECO approach minimized the importance of transmission congestion and rejected the locational pricing model as too complicated and unnecessary. Instead, the PECO model would treat the entire PJM system as a single zone.

In essence, the PECO model priced all transactions through the spot-market at the "unconstrained" price, based on a hypothetical dispatch. To the extent that the actual dispatch encountered transmission constraints, the PECO model would pay the more expensive generators to run and average these congestion costs over all users.

The model included two other notable features. First, in the face of transmission congestion, the generators that were constrained not to run would be paid nothing, even though they had bids below the "unconstrained" price. There was objection to adopting any system that depended on paying generators not to run, with the attendant discrimination and perverse incentive effects. Second, market participants had the option to schedule bilateral transactions separate from the bid-based economic dispatch of the ISO, with a separate payment for their share of the total congestion cost. This flexibility to use bilateral transactions or to participate in the coordinated spot market was a major design objective not to be abandoned.

This pricing system is representative of a zonal approach, and has much in common with zonal systems adopted elsewhere in the world.[52] However, should the system become constrained, the two exceptional features noted above would create a powerful and perverse

---

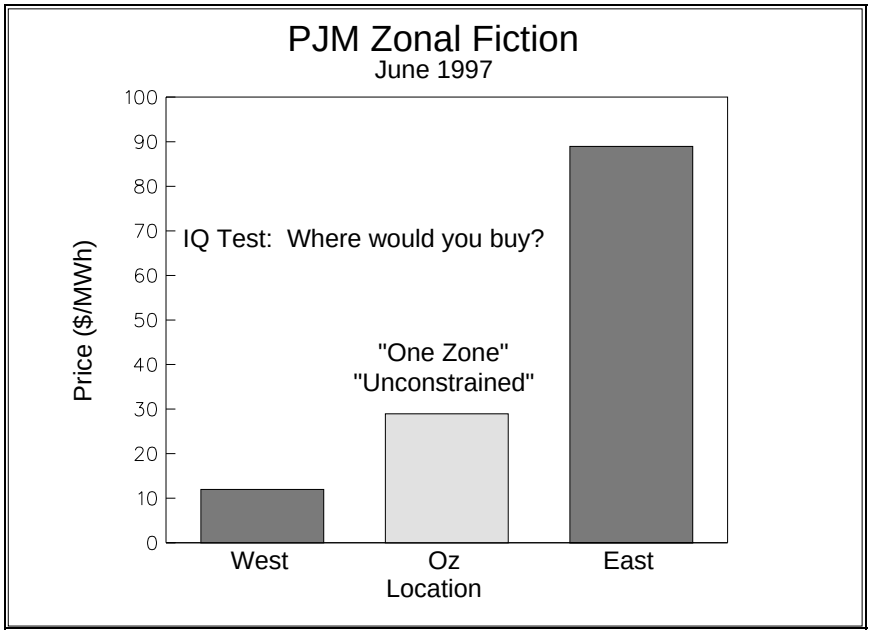
Journal, Vol. 10, No. 1, January/February 1997, pp. 24-31.

[51] For details, see William W. Hogan, "FERC Policy on Independent System Operators: Supplemental Comments," Federal Energy Regulatory Commission, Docket No. PL98-5-000, Washington DC, May 1, 1998.

[52] Here the issue is pricing for transmission congestion. The recovery of embedded costs of transmission investment through access charges is a separate matter that is amenable to zonal approach.

incentive. If there were no transmission constraints, there would be no transmission congestion and everything would work as with the locational pricing system. But when congestion appeared, everything would be different. The supporters of the zonal approach argued that the total cost of congestion would be small, summed over the year, and therefore any inefficiencies could be safely ignored.

Ignoring a difference between prices and marginal costs is a safe practice in a regulated world without flexibility and choice. The incentives don't matter and the small costs get lost in the larger system. It can work inside the closed black box. But the cost of ignoring a gap between prices and marginal costs in the world of choice can be large indeed. Witness the events when the PJM system became constrained, starting in June of 1997.



The data for a representative constrained dispatch found the marginal cost in eastern PJM at about $89 per MWh, when at the same time the marginal cost in the west was $12 per Mwh. At the same time, the "unconstrained" price for the "One Zone" (Oz) was approximately $29 per MWh. The incentives were clear. A customer could buy from the spot-market dispatch at $29, or it could arrange a bilateral transaction with a constrained-off generator in the west at a price closer to $12.[53] The small average congestion cost would be the same either way, and would not affect the choice. The choice, therefore, presented a low-level IQ test.

Faced with these incentives, constrained-off generators passed the IQ test. They quickly arranged bilateral transactions and scheduled their power for delivery, thereby exceeding the transmission limits. This, in turn, required the ISO to constrain the output from some other generator, who would then follow the same direct path to a bilateral schedule rather than sit idle and collect nothing. Soon the ISO had no more controllable generating units with which to manage the transmission constraints. Unable to fix the perverse pricing incentives, the ISO resorted to administrative mechanisms to prohibit bilateral transactions or declare a "minimum"

---

[53] Power Markets Week, September 1, 1997, p. 13.

generation emergency during the peak generation period. In effect, while restructuring to facilitate a market, the unintended consequences of superficially simple pricing spawned administrative rules to prohibit the market from responding to the price incentives when they mattered most. Shackled with inconsistent pricing rules, the ISO had to resort to direct preemption of market choices.

The point was made in a dramatic way. The important issue is not the total cost of congestion, which may be small on average. The point is the incentives at the margin when the system is constrained. In designing the rules for transmission pricing and access for a competitive market, it matters little what the rules are for periods when the system is unconstrained. The important question is how the rules deal with the market when the system is constrained. Even if the total cost of congestion might be modest over the year, the gap between $29 and $12, or $89 and $12, is more than sufficient to get the attention of market participants. Given the margins in this business, they will change their behavior for $1. And the changes in behavior can substantially affect system operations; in fact, the whole point of electricity restructuring is that changes in behavior can affect system operations and lead to different patterns of electricity use and investment.

In the locational pricing system, the perverse incentives would not arise. Given the same facts as above, the locational prices would equal the marginal costs. Those customers purchasing power from the spot market in the east would have seen $89 as the price. True, they could have arranged a bilateral transaction with a generator in the west, paying $12 for the energy. But they would then face a transmission charge of ($77=$89-$12), making them indifferent at the margin, just as intended. Likewise, customers in the west would pay $12 and have no incentive to change. Every generator would be producing at its short-run profit maximizing output, given the prices. The market equilibrium would support the necessary dispatch in the presence of the transmission constraints. Spot-market transactions and bilateral schedules would be compatible. Flexibility would be allowed and reliability maintained consistent with the choices of the market participants.

The PJM ISO was fully aware of the perverse incentives of zonal congestion pricing and the problems they created, but without the authority to change the pricing rules it had no alternative but to restrict the market. Faced with this reality, the FERC acted to approve the locational pricing system that became operational in PJM at the beginning of April of 1998. The developing experience should be better understood to avoid the pitfalls of the complicated zonal "simplification."

The most obvious flaw in the zonal argument is in its very definition. If the nodal prices are not materially different, then there is no need to aggregate into zones. The nodal prices would already be simple to use in the market. Apparently the move to aggregate nodes into zones is really an effort to treat fundamentally different locations as though they were the same.

If market participants had no market choices, then there would not be much effect of

such zonal aggregation, other than a certain amount of cost shifting. But a central objective of market restructuring is to give market participants as many choices as possible. Further, we expect that market participants will respond to profit incentives. If we don't "get the prices right," the market actors will respond to the prices and make choices that at best would significantly raise costs and at worst would dangerously compromise reliability.

There are many ways that things can go wrong. The PJM 1997 experiment with a zonal pricing system collapsed as soon as the system became constrained. New England faced the problem of zonal incentives to build new generation in locations that would exacerbate constraints. In Australia, a zonal pricing system has complicated the ability to offer transmission rights that match the real capability of the system. The same physical laws that govern nodal pricing make it impossible to define a complete set of zonal transmission rights, or to guarantee that generators can always participate in the market. By contrast, with a nodal pricing system such as in New Zealand, point-to-point transmission contracts could be defined in a natural way that is inherently consistent with the pricing regime and the real capacity of the grid. Just such a system of transmission contracts is operational with nodal pricing in PJM, and plays a central role in the recently FERC-approved nodal pricing system for the New York ISO.[54]

Similar experiences revealing the hidden complications of zonal pricing can be found from England to California, with different ad hoc rules applied to create more and more complex structures to fight against the choices of market participants confronted with administrative prices. Often it is hard to recognize the connections among the isolated ad hoc decisions, or to see the root of the problem in offering choices without getting the prices right. New England is not alone, but it could learn from the mistakes of others.

A great advantage of a nodal pricing system is that it creates incentives that are "self-policing."[55] Competitive market generators and loads could bid into a spot market and find that the economic dispatch result created a solution that would meet the no arbitrage condition. The attendant locational prices would be such that, as shown in the figure, every generator with a bid less than the price at its location would be running, and generators who had bid more than the market clearing price would not be running. There would be no artificial incentive to deviate from the market equilibrium solution.

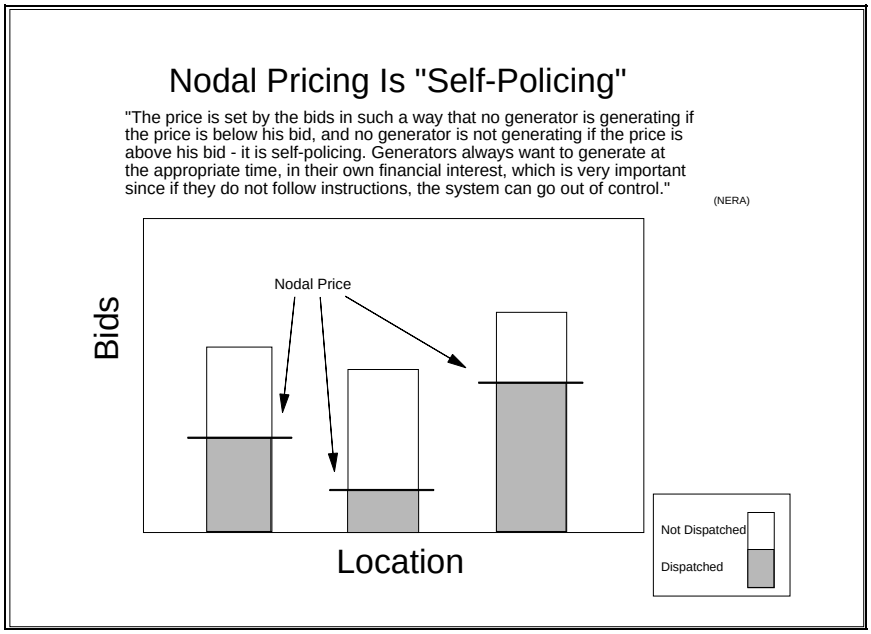
By contrast, a zonal pricing system must by definition create conflicting incentives. Set aside the complications about how to determine the zonal price. Whatever the rule, the zone will by definition have a single price. For generators who have bid less than the zonal price but who cannot run because of transmission constraints, there will be a strong incentive to leave the spot market and schedule a bilateral transaction, just as in the PJM experience. By contrast, for

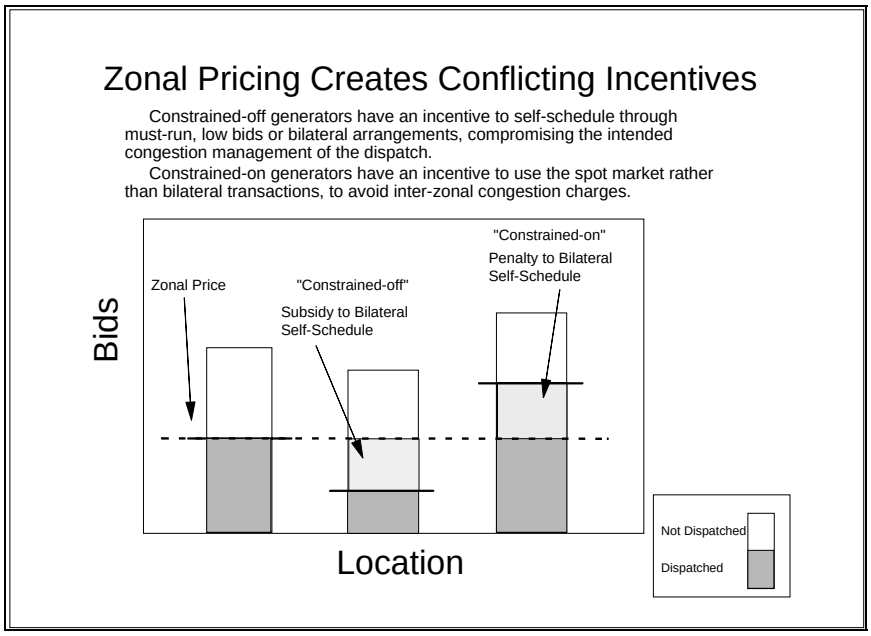
---

[54] The New York system proposed an aggregation for final loads, at least as a transition mechanism to deal with metering problems. The FERC explicitly objected to this proposed zonal aggregation for loads, calling for any metering changes needed to apply a full nodal pricing system.

[55] The term comes from National Economic Research Associates (NERA).

generators who have bid more than the zonal price but must run, again due to transmission constraints, there is a bias in favor of the spot market and these generators could not participate in the bilateral market. Furthermore, in a multi-settlement system these constrained-on generators would have a strong incentive to generate below their commitment and make up the imbalances at the lower zonal price. The details would depend on the particular rules and market setting, but the general conclusion holds that the incentives created by aggregation would greatly complicate operation of the market.





The real impact of zonal pricing is to create more administrative rules, poorer incentives for investment, demands to pay generators not to generate power, and proposals to "socialize" the higher costs by using the taxing power of the ISO. This is not the way of a market. It creates more problems than it solves.

Furthermore, there really are no major complications in implementing full locational pricing, once you look closely at what is required. The fears are misplaced, and the benefits are real and substantial.

Consider some of the arguments:

Is transmission congestion a small problem? No. On close inspection, few systems are really unconstrained. And when the constraints bind, the effects can be very important. Transmission congestion costs can easily exceed generation costs at the margin. The incentive effects have major commercial implications. Before the fact, zonal advocates argued that PJM would consist of at most only a few zones and constraints would be rare. In the event, the PJM market environment saw significant constraints in 15-20% of the hours in the first six months of operation, often under conditions that would be important in capturing commercial profits. And the number of zones needed to respect the commercially significant price differences in PJM has been far more than the promised few. Zonal pricing is not simple.

Are nodal prices produced by a black box? No. Given the dispatch, the prices are easy to calculate, explain and audit. There is ample operational experience to dispel the notion that nodal pricing is too hard. The engineers know how to do it, and have been doing it for years. The nodal prices have always been there; we just haven't used them for market transactions.

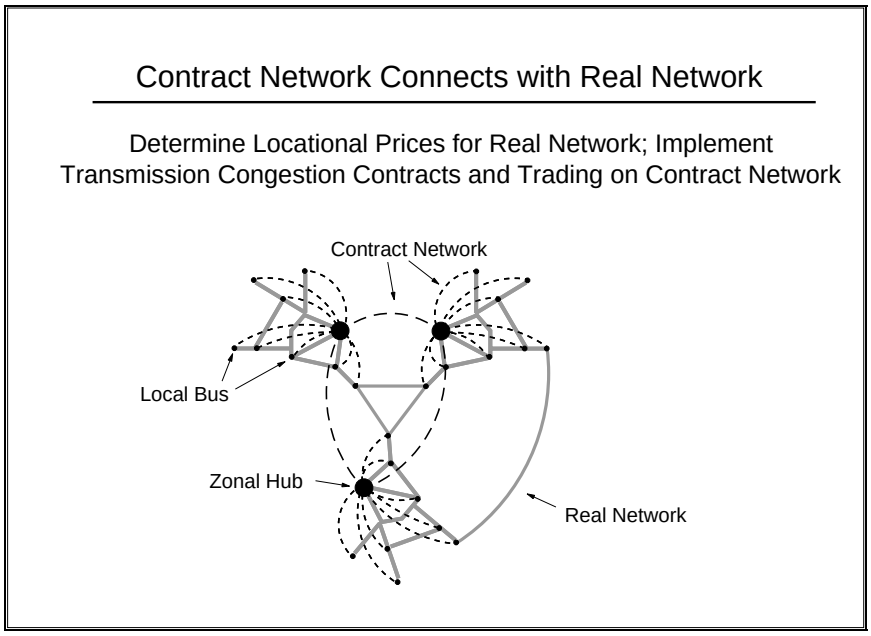
Doesn't nodal pricing preclude transmission price certainty? No. To be sure, we do not know in advance what the spot price will be, just like in any market. But those who want transmission price certainty can acquire transmission congestion contracts. And those who do not want to pay in advance for price certainty, and want to rely on the spot market, cannot socialize the costs by making others pay for the congestion they create.

The list of misconceptions about the pricing debate is longer. Given the fundamental underlying differences in marginal costs, it is not so easy to define the zonal price. It is not an easy matter to set or later change the zonal boundaries. The inherent averaging of zonal prices tends to remove incentives for energy efficiency or distributed generation. And so on. Perhaps the most oft-repeated point of confusion has to do with the impact of zonal aggregation on the ability to exercise market power.

Won't zonal aggregation mitigate market power? No. Real elimination of the physical constraints would help reduce market power where it exists. But administrative aggregation into zones simultaneously increases and obscures market power. Under the zonal approach, favored generators could take advantage of the real physical constraints, but their higher charges would be socialized and averaged over all system users, hidden from view. Market power can be a problem, but the problem is neither created by locational pricing nor resolved by zonal aggregation.

More recently, there has been the argument that the market needs zonal aggregation to support simplified trading. Won’t nodal pricing destroy market liquidity? Apparently not. The issue is more complicated in the case of electricity than other markets because the open access spot market creates its own form of liquidity that may obviate the need for vigorous trading of bilateral contracts. However, even for trading in contracts, this is largely an empirical

question. The early returns from PJM suggest that the predictions of no liquidity in the market have been quite wrong at the western hub. Of course, no market has complete liquidity at every location. Typically there are trading hubs and the liquidity is found at the trading hubs, not at every location. This use of trading hubs is valuable and fully functional under nodal pricing.



Isn't a simpler system possible? Yes. The nodal pricing approach is completely consistent with a hub-and-spoke description of the market. One or more trading hubs can be established. The transmission charge for moving along the spokes, to and from the hub, is just the difference in the locational price and the hub price. This would be in contrast to a price of zero along the spoke implicit in a zonal aggregation. A hub can be selected as a single location or as a fixed portfolio of locations. This approach captures most of the intended simplification of the zonal model without embracing the hidden defects of aggregation. There is no mystery here. The hub-and-spoke approach is the system now working in PJM.

There is nothing unusual in nodal pricing. It is the natural system that falls out of an analysis of competitive market marginal-cost pricing principles in the context of the physics of the electric network. Nodal pricing does not solve all problems in electric market design, but it turns out to be important in dealing with some of the otherwise most intractable problems created by the special nature of the electric grid and the complex network interactions. Furthermore, if despite all the evidence zonal aggregation is commercially attractive, it presents a business opportunity for its advocates and need not be imposed by the ISO. But practical experience and theoretical analysis both support the conclusion that for the independent system operator, nodal pricing is the simplest system that actually works in the context of a market with choices and flexibility.

Get the prices right, and it is much easier to rely on the market.

--end--