**Statement of**

**William W. Hogan**[1]

**Regarding the NEPOOL Proposal for a Congestion Management System**

**FERC Docket no. ER99-2335-000**

April 20, 1999

## Introduction

On March 31, 1999, the New England Power Pool (NEPOOL) submitted to the Federal Energy Regulatory Commission (FERC) a "Congestion Management System Proposal for NEPOOL." The purpose of the present statement is to comment on the substance of the NEPOOL proposal and place it in context in terms of support of a competitive electricity market and previous FERC decisions. In summary, the current proposal replaces and substantially improves upon the earlier design for congestion management in NEPOOL. The new locational pricing system is consistent with a competitive electricity market, will support the functions of the New England Independent System Operator (ISO-NE), is consistent with the previously FERC approved congestion management systems for the Pennsylvania-New Jersey-Maryland Interconnection (PJM) and the New York ISO (NYISO), and supports a workable system of point-to-point transmission rights in the form of financial congestion rights (FCRs).

Following the FERC decisions in late 1998, which gave NEPOOL the opportunity to revise its procedures for transmission congestion management and connection of new generators, along with many others I joined in the discussions of the Congestion Management System (CMS) subcommittee, working initially with the Conservation Law Foundation, Westbrook Power and New England Power Company. This group expanded to what eventually became a majority in support of the view that a nodal pricing system for day-ahead and real-time operations would be the simplest and most efficient means of managing system congestion under the direction of the New England ISO (ISO-NE).

Based on these discussions and my analysis of the underlying economics of transmission usage in the context of a competitive electricity market, the present comments address the related issues of locational pricing and financial transmission

---

[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).

rights, the main pillars of the proposal.[2] In addition, the final section considers the connection with the companion proposal for a Multi-Settlement System (MSS).

**Locational Prices For Transmission Use**

The problems associated with pricing and incentives under the earlier NEPOOL congestion management system were virtually identical to the problems of the nearly identical pricing system applied in PJM in 1997. That system failed because it presented the wrong incentives to market participants. In 1998, PJM implemented a nodal pricing system to provide incentives that would be consistent with operating a reliable competitive electricity market. This fully locational pricing system in PJM recently observed its first anniversary of successful operation. In an attachment, I summarize this history and the evidence that, by analogy, suggests that the current NEPOOL CMS proposal will work well and will support FERC's objectives.[3]

A central and pervasive problem encountered in managing transmission congestion usage arises because of the complications of loop flow. To recall the issues, consider the market equilibrium in an electricity system. For simplicity, ignore the effect of transmission losses. When the system is not constrained, the force of competition would tend to eliminate any price differences across the system. In equilibrium, the unconstrained condition would imply a single market-clearing price for electricity at that time. This is a familiar result and a guiding principle that tends to appear in virtually all pricing systems for spot-market transactions. The common sense and intuition leading to this conclusion are supported by simple analysis and broad experience.

However, when the transmission system is constrained, experience elsewhere and simple analyses of radial connections, can mislead. It is common to argue that a single transmission constraint would separate the system into two markets or zones, with two prices in equilibrium. However, this common argument is wrong, often seriously wrong, in an interconnected grid with loops. In fact, in a real transmission network it is theoretically possible that a single transmission constraint can produce different prices at every location in the system. This follows because the loop-flow effects include the simple fact that every location has a different impact on the constraint. The theoretical argument has been supported, strongly supported, by the experience in PJM, which is the only publicly available source of data about market-based nodal prices.

The details are discussed further in an appendix, which summarizes the underlying economics of a competitive electricity market. The basic implication is in the principle of getting the prices right. If we wish to give market participants a great deal of flexibility in scheduling bilateral transactions, buying and selling through the spot market, and minimizing the use of administrative controls, then it is important to present

---

[2] Note that the focus here is in transmission usage and pricing to support congestion management. This is separate from the method used to pay for the embedded cost of the transmission system, where we would be dealing primarily with sunk costs and the "license plate" approach would be available.

[3] William W. Hogan, "Getting the Prices Right in PJM: Analysis and Summary: April 1998 through March 1999, The First Anniversary of Full Locational Pricing," Harvard University, Center for Business and Government, April 2, 1999.

them with market price incentives that are consistent with competition subject to the reliability constraints on the grid. This means that prices can be different at every node. Transactions through the spot market should be at the nodal prices. And for bilateral schedules, the transmission usage charges should be equal to the difference of the nodal prices at the source and destination.

With this set of pricing incentives, there would no longer be any need to place restrictions on connection of new generators because of their impact on congestion. The congestion prices would incorporate the right incentives in this regard, and the only connection charges needed would be those to complete the local attachment to the grid.

The principal alternative to the nodal pricing approach is to aggregate many locations into one or a few zones. The previous NEPOOL proposal was to treat all of the NEPOOL as a single zone. This would not work, but the argument remained that it might be possible to aggregate the system into a few zones. The accompanying paper examines this argument in further detail, and considers the implications for the same argument in PJM, where the experience is that a system with a few zones would create added complications compared to the simple system of using nodal pricing.[4] The fundamental problem with the zonal approach for congestion management is that it works well only under conditions where it is not needed.

The usual claim in zonal aggregation is that the zones will be created to aggregate nodes that have the same impacts on the constraints. However, if there are such collections of nodes, then the result of nodal pricing would be that the nodes would have the same prices. Hence, under this circumstance there is no advantage to zonal pricing compared to nodal pricing because the prices would already be the same. And if the zonal assumption is violated, so that the aggregation applies to nodes with different prices, then all the incentive problems arise within the zone that created the difficulties that we saw in PJM in 1997. Either we need so many zones so that there is no simplification compared to nodal pricing, or we create perverse pricing incentives that lead to administrative controls and barriers to entry that compromise the fundamental objective of supporting a competitive market. Despite claims to the contrary that the details don't matter, there is no zonal pricing system in the world that confronts transmission congestion without producing the perverse incentives and administrative interventions that are avoided with a nodal pricing congestion management system. Nodal pricing is the truly simple approach.

The NEPOOL CMS proposal is at its core a nodal pricing approach. For generators, it is a fully nodal pricing approach, as in PJM and New York. For load, the NEPOOL proposal includes a zonal aggregation of prices. This is similar to the aggregation for load in New York, where the argument was based on a temporary problem with metering load busses. In New England, the motivation for this zonal aggregation is that some customers in New England have "standard offer" contracts that were defined under the old congestion management system. There is a fear that fully

---

[4] See the appendix in William W. Hogan, "Getting the Prices Right in PJM: Analysis and Summary: April 1998, through March 1999, The First Anniversary of Full Locational Pricing," Harvard University, Center for Business and Government, April 2, 1999.

locational pricing for load might present opportunities for gaming these fixed price contracts in ways that were not anticipated and not consistent with the regulatory intent. Furthermore, there is a general concern with immediately presenting final customers with different prices when congestion appears.

Both of these motivations would support a transition period with aggregation of prices for final loads. In the short-run, the perverse incentives that appear with zonal aggregation are more severe for generators than for loads. However, in the long-run, the intended increased use of demand-side bidding and load management techniques would argue for movement to a full nodal pricing system for both generation and load.

Note that the NEPOOL description of the load-zone aggregation criterion confronts the potential contradiction identified above. The current document states that the load "[z]ones will be established to group buses that have the same or nearly the same impacts on potentially limiting interfaces when energy is injected into the system.."[5] However, if this were true, then the nodal prices would also be the same or nearly the same. Either this means that aggregation is not necessary for load, or that there will be many zones. If the latter, then the zones might be unlikely to match the regional coverage of the standard-offer contracts. Since this would be inconsistent with the original purpose of the load aggregation, my expectation is that as the mechanism is developed further, there will be aggregation of different prices, reinforcing the need to make the load-zone approach temporary and move to a fully nodal system.

The use of load zones raises technical details that would be important. For example, bilateral schedules to deliver power from one location to another should be treated under the nodal price system. A bilateral transaction from node A to node B should be charged the difference in congestion costs between node A and node B, rather than the difference from node A to the zone containing B. Only when the final load actually withdraws power from the system should there be a credit to bring the load price to the average for the zone. This would avoid, for example, the problem of a bilateral schedule from A to B, matched with a bilateral schedule from B to C, receiving the load credit twice.

Finally, the NEPOOL CMS proposal anticipates and is fully consistent with the creation of one or more trading hubs, as in PJM. This approach has all of the advantages of a zonal "simplification" except that transmission to and from the hub is not free. The hubs can be defined as a fixed-weight portfolio of nodes, and the price for moving between any location and the hub is just the difference in the nodal price and the hub price. As discussed in the appendix, this allows for simplified trading and is fully consistent with the nodal pricing approach. For example, the western hub in PJM has become a location for a highly liquid market and a newly introduced futures contract.

**Financial Congestion Rights**

If spot market transactions are priced at spot market prices, then the congestion price of transmission can change from hour to hour, and can be quite volatile. This

---

[5] NEPOOL CMS Proposal, March 31, 1999, p. 7.

volatility creates risks for long-term transactions, and there is a natural interest in creating transmission rights that would allow market participants to fix the cost of transmission usage in advance.

The usual first proposal is to create some form of firm transmission or physical rights to the deliver power. However, this approach is confounded by the complications of loop flow. For essentially the same reasons that loop flow creates different prices at every location in a constrained system, physical rights would imply great complexity in actually using the full capacity of the transmission system. Physical rights or firm transmission in the physical sense is not really attainable, at least in this competitive market environment.

In a context of a locational pricing system, however, there is an alternative that is superior to physical rights. The alternative is to define a set of financial rights for the congestion costs between locations.[6] The Financial Congestion Right (FCR) is simply a contract to collect from ISO-NE the difference in congestion costs at two locations times the amount nominated in the contract. This contract is the same as the "Fixed Transmission Rights" (FTRs) in PJM, or "Transmission Congestion Contracts" (TCCs) in the New York proposal. Hence, an FCR for 100 MW from A to B would entitle the holder to the difference in congestion costs at A and B times 100 MW.

If a market participant wants to ship 100 MW from A to B, then the congestion charge for the actual use of the transmission system is exactly equal to the payment under the FCR, and the use is as if there were physical rights to this transmission. However, the efficacy of the FCR does not depend in any way on actually matching the contract and actual use. In effect, the FCR is superior to physical rights, or can be thought of as a perfectly tradeable physical right. Just as nodal pricing incorporates all the effects of loop flow and simplifies the operation of the market, the FCR incorporates the same impact of loop flow and provides the economic equivalent of the physical rights that we don't know how to implement.

The FCR, therefore, allows market participants to determine the price of transmission in advance. Those who are willing to pay for the FCR in advance can know in advance what it will cost to use the transmission system. Of course, those who are not willing to pay in advance, and prefer to use the spot market, cannot be guaranteed in advance a particular spot market price.

If the total allocation of FCRs is simultaneously feasible, then the revenues collected by the ISO through the nodal congestion prices will be sufficient to cover the payments of the FCR obligations, hour by hour, as long as the corresponding capacity of the grid is available. There may be excess revenues, and presumably these revenues would reduce the cost of transmission fixed charges, as elsewhere. If the full allocated capacity of the grid is not available, either due to outages or external loop flows, the congestion revenue deficit in some hours would be balanced by the surplus in other hours, with a monthly true-up applied pro-rata if there is an aggregate deficiency.

---

[6] In principle, the price for losses could be included, but we focus here on congestion for simplicity.

Consistent with the nodal pricing system, all FCRs should be defined from point to point. This is true even when the FCRs are used by load in a zone. The point-to-point definition provides the correct hedge, and allows for easy transition when and if the definition of the load zones change. As anticipated in the NEPOOL proposal, FCRs could include any nodes or portfolio of nodes such as the trading hubs.

Under the CMS proposal the allocation of FCRs for the existing grid would be assigned to transmission customers, who would in turn allocate the benefits to final customers. Since the existing grid is in place, and the FCRs should be fully tradeable, this allocation should have no impact on the going-forward incentives or efficiency of the market. This allocation of the initial rights is consistent with the policy in PJM and in the New York proposal. There will be debate on this matter from the minority perspective on the CMS subcommittee that some or all of the FCRs for the existing system should be allocated to generators. This is primarily an equity or legal question, but the debate should have little to do with the economic incentives that will govern future use of the the system.

To the extent that there are additional possible FCRs beyond those allocated to load, there could be an auction of the type adopted in PJM. On April 15, 1999 PJM began auctioning FTRs that would be simultaneously feasible along with the other FTRs already allocated. The initial auction is for the month of May, peak and off-peak. The PJM FTR auction allows bidders to specify any desired portfolio of FTRs and awards the most valuable combination subject to the limits of simultaneous feasibility of all FTRs. Just as with nodal pricing, the price of the winning awards is the market-clearing price, which should greatly simplify the preparation of bids.[7] The auction revenues serve to reduce transmission fixed charges. A similar auction proposal appears in the NEPOOL CMS proposal.

As for future investment and the allocation of incremental FCRs for transmission expansion, there would be an important incentive effect. As discussed in the appendix, acquiring the incremental FCRs would be part of the incentive for making transmission investments, and these should be allocated to those who make the investment. This principle is embodied in the NEPOOL CMS proposal, although the details have not been worked out. The problem of defining and balancing the incentives for transmission investment is difficult, and there may be no completely market-based approach that would deal with all the externalities. However, it is probably true that any market-based system would have to allocate incremental FCRs to those making transmission investment, and the NEPOOL CMS proposal adopts the right principle in this regard.

**Multi-Settlement System**

There is a separate proposal before the FERC for NEPOOOL to implement a day-ahead market and a multi-settlement system. There is no controversy here, and my understanding that the multi-settlement system has unanimous support among the

[7] For further details, see the PJM web site at www.pjm.com under "PJM Training."

NEPOOL participants. Without elaborating fully on the multi-settlement system, one comment is in order as to the connection to the CMS proposal.

The essence of the multi-settlement system is that day-ahead schedules and bids are resolved through the bid-based, security-constrained, economic dispatch to produce day-ahead contracts for delivery of energy, reserves and so on. Part of this day-ahead schedule includes the day-ahead locational prices that are applied using the same principles as outlined above. All the contracts and schedules are settled at these day-ahead prices. This is the first settlement.

Then in real time, adjusted bilateral schedules and spot market bids produce the actual use of the system and real-time locational prices. In real time, the deviations from the day-ahead contracts and schedules are settled at the real-time prices. In this sense, the day-ahead contracts are seen as financial contracts that do not require physical delivery, as long as the real-time settlement, the second settlement, is at the real-time prices.

This two-settlement system could be extended to any number of settlements, such as an hour-ahead, but the principles would be the same. Similar proposals for multi-settlement systems are under development in PJM and part of the design in New York. There are many potential advantages to such multi-settlement systems, particularly in expanding the ability of flexible demand to bid in and fix prices a day-ahead when there would be more opportunities to adjust loads.

The main principle to emphasize for the multi-settlement is to maintain the consistency of the settlements and the prices. It is possible to get this wrong, as in the first implementation of the market in England & Wales when they fixed the prices a day-ahead but then allowed adjustment in the actual generation inputs. Albeit inadvertent, this created perverse incentives that had to be dealt with through administrative restrictions. But with the prices consistent, day-ahead prices for day-ahead quantities and real-time prices for real-time quantities, the perverse incentives should not arise.

In the case of FCRs, application of this principle means that FCRs should be settled by the ISO-NE in the first settlement at the day-ahead prices. It is my understanding that this is anticipated in the NEPOOL multi-settlement proposal. In effect, the FCRs define one-possible use of the system, and the day-ahead contracts define another. Only one internally consistent set can be in place at any time.

This does not mean that market participants cannot use FCRs to hedge against real-time prices. They would merely need to schedule their FCR in the day-ahead market in order to accomplish this objective. But the ISO cannot simultaneously determine a day-ahead dispatch and settle FCRs at the real-time prices. This same principle is recognized in the New York and PJM proposals.

**Conclusion**

The NEPOOL CMS proposal represents a major improvement over the previous system, which would never have lasted and would have created significant problems in the face of substantial transmission congestion. Nodal pricing is the truly simple approach that is consistent with a competitive electricity market. As a transition, limited

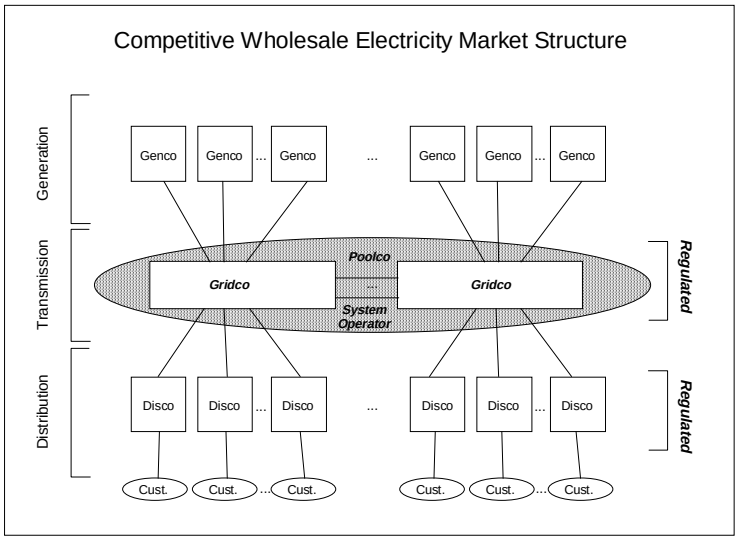
zonal aggregation for loads would solve special problems in New England. The complementary system of FCRs gives the only known workable mechanism for providing the equivalent of firm transmission rights without encountering the substantial problems that physical rights would entail. The NEPOOL CMS proposal is consistent with the similar approaches taken in PJM and New York, which should help with coordination in the eastern market. The FERC has approved the essential details in these other systems, for good reason. The same approval should be applied here with the usual guidance to develop the details further in the same direction to support open access and non-discrimination for a competitive electricity market.

## Appendix
## 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.[8] Here we focus on the implications 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

---

[8] 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.

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 prisoners 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.[9] 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. 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.

[9] 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.