Transmission Pricing and Congestion Management: Efficiency, Simplicity and Open Access

Shmuel S. Oren
University of California at Berkeley
Berkeley, CA 94720
TL. (510) 642-1836 FAX (510) 642-1403 Email OREN@IEOR.Berkeley.edu

Abstract

Transmission pricing and congestion management are the key elements of a competitive electricity market based on direct access. They have also been the focus of much of the debate concerning alternative approaches to the market design and the implementation of a common carrier electricity system. This paper focuses on the tradeoffs between simplicity and economic efficiency in meeting the objectives of a transmission pricing and congestion management scheme. I contrast two extreme approaches: the postage stamp approach vs. nodal pricing. The paper questions the wisdom of the nodal pricing paradigm on the grounds of its rigidity and complexity. I argue that the theoretical efficiency properties of nodal pricing are unrealistic and do not justify the implementation drawbacks of the approach. The paper explains the underlying principles of least cost congestion relief, adopted in California that treat congestion relief as an ancillary service and enables the ISO to relieve congestion efficiently with minimal intervention in the energy market. I also discuss zonal aggregation and describe a new zonal priority network access pricing that complements interzonal congestion pricing by offering a market mechanism to guide intrazonal congestion management and provide economic signals for location of generation resources.

Introduction:

There is general agreement among academics practitioners and policy makers that direct access to the transmission grid is the essential centerpiece for a competitive electricity market. Order 888 and Order 889 of the Federal Regulatory Energy Commission (FERC) reflect the role of direct access as the foundation for the electric power industry restructuring. These orders provide guidelines for nondiscriminatory transmission pricing and mandate timely disclosure of available transmission capacity but do not prescribe a particular approach to the institution of direct access. However, the prevailing restructuring paradigm being adopted in many states in the US has two key features: functional unbundling of generation transmission and distribution and the transfer of control over the transmission system to an Independent System Operator (ISO). The establishment of the ISO as a key institution in the emerging competitive electricity markets is based on the consensus that the physical characteristics of electricity impose requirements for real time central coordination in order to assure reliable service. However, the extent of centralized control and “market management” that is needed to assure system reliability and that is desirable from a social efficiency perspective has been a subject of public debate. That debate has polarized the restructuring approaches adopted so far on the east and west coasts of the US. This divergence manifests itself in the transmission pricing and congestion management protocols the centerpieces of a direct access system and two of the key functions of the ISO.

In an open access, competitive electricity system a transmission pricing scheme should fulfill several functions and meet various criteria:

- Generate revenues to compensate the owners of transmission assets
- Produce economic signals for efficient rationing of scarce transmission resources
- Produce economic signals for efficient investment in transmission and for efficient location of new generation capacity and loads
- Be simple to implement, transparent and conducive to energy trading.

Like most lists of desiderata, the above requirements can only be met partially by any practical transmission-pricing scheme, which requires some tradeoffs and compromises. One of the basic tradeoffs in this context is between simplicity and short-term efficiency. The questions that one must address in making this tradeoff are:

- How precise need the short-term economic signals be to move the system toward long-term efficiency?
- What is the economic cost of a simpler and less accurate pricing scheme?
- What is the correct level of precision in short term price signals given the approximations in the system models, the formulation of objectives and the available computation technology?
- How we compare the economic value of an accurate ex-post price signal to an approximate ex-ante price signal?

Another important tradeoff concerns the desired level of decentralization. The two basic paradigms for optimal resource allocation are the central planning approach versus the decentralized "invisible hand" approach. The equivalence of the results in an idealized theoretical setting has enabled economists to simulate market outcomes by using optimization models. Indeed market simulation is the proper use of optimization technology in a competitive market setting. The endlessly debated philosophical question is to what extent should a market simulation model be used to manage the market. This debate dates back to the 1920's when the concept of planned economies was first proposed. Using a market simulation model to set prices as advocated by some in the electricity restructuring debate is somewhat analogous to using statistical poling to determine (rather than just forecast) election results. The physical characteristics of electricity and the network aspects of electricity transmission justify a certain degree of central coordination in order to maintain system reliability. However, the extent to which central intervention based on optimization models should be used to insure short-term economic efficiency is a debatable policy decision with direct consequences for transmission pricing and congestion management protocols. The argument for a centralized market management approach is that given the need for central coordination (for reliability considerations) it would be foolish not to take the extra step and optimize resource use so as to maximize short term efficiency. The counter argument is that, multiperiod efficiency need not collapse to the sum of single period efficiencies. Massive integration of systems on cost-based efficiency grounds can come at a loss of gains from competition and innovation born of profits. Decentralization and minimal central intervention in the market will promote long term efficiency by facilitating interaction between buyers and seller, customer choice, intermediation and technological innovation. The basic question is then does the long term potential gain justify the short term losses caused by the response lags and inefficiencies of a decentralized system.

The differences among the various proposed schemes for definition of transmission rights, transmission pricing and congestion management can be categorized along several dimensions as follows:

- Physical vs. financial transmission rights
- Link based vs. node based (point to point) definition of transmission rights
- Access based pricing vs. usage based pricing
- Locational differentiation in tariffs: nodes, zones or uniform prices
- Ex-ante vs. ex-post pricing
- Bundling of transmission service and energy vs. treating energy and transmission service as separate commodities
- Congestion management through efficient generation dispatch vs. efficient congestion relief.

It is beyond the scope of this paper to provide a comprehensive survey of all the proposed approaches and even if we tried we would probably miss some. Instead we will focus on two basic ideas related to the simplification and decentralization of decision making in the deregulated electric power industry. First I will explain how the California congestion management approach has been able to separate the energy market from the transmission market by, effectively, treating congestion relief as an ancillary service. Next I will discuss the issue of zonal aggregation and describe a new priority zonal network access pricing approach that may be viewed as an extension of the familiar postage stamp approach. The proposed scheme enables efficient intrazonal congestion management based on a relatively simple ex-ante transmission tariff. In order to put these ideas in context I will first contrast the two extreme approaches on the simplicity vs. efficiency tradeoffs.

Simplicity vs. Efficiency: Is Nodal Pricing Worth the Trouble?

Two opposite extremes in terms of the tradeoff between short-term efficiency and simplicity in transmission pricing are the nodal pricing approach and the postage stamp approach. In the latter transmission pricing takes the form of a fixed ex-ante charge per MWh for transmission service between any two points in the grid. The simplicity and certainty of this approach is compelling from the point of view of energy trading over the grid. However, it has been argued that the lack of locational differentiation results in no economic signals to investors and users for efficient location of new load (e.g. production facilities) and for the location of new generation and transmission lines. Furthermore, postage stamp transmission pricing does not elicit economic signals from customers that could be use to manage congestion efficiently. There is, however, little evidence as to the magnitude of the efficiency losses resulting from the lack of correct economic signals and the debate is raging with regard to how precise these signals need be to recapture most of these losses.

Motivated by short-run efficiency considerations, the nodal pricing approach advocated by Hogan [1992] manages congestion and sets transmission prices through a centralized energy market based on economic dispatch. The basic idea of the nodal pricing approach is to organize the market as a pool in which generators (and ideally loads) submit hourly bids for node specific injection and withdrawals of power to an Independent System Operator (ISO) with full coordination and price setting authority. The ISO minimizes the total system's gain from trade (demand bids less supply bids) subject to transmission and reliability constraints. The price at each node is then set to the incremental bid price of the most expansive unit generated or consumed at the that node. These nodal prices become the hourly prices charged to loads and paid to generators at the respective nodes. When there is no congestion all nodal prices are in theory identical. However, even congestion on a single link could result in a different price at every node in the system (in the WSCC there are around 2500 such nodes).

Proponents of the nodal pricing approach claim that bilateral transactions can be readily accommodated within this framework. A physical bilateral transaction can be scheduled as if the injection submitted a zero bid and the load submitted an infinite bid. Such a transaction is then subject to an ex-post transmission charge that equal the opportunity cost of the transaction, i.e., the cost difference of selling the power to the pool at the injection node price and buying it back at the withdrawal node price. Thus, the transmission charge between any pair of nodes is set ex-post to the nodal price difference between he nodes. The cost off transmission, therefore, varies between each pair of locations and is only known to energy traders after the fact. Bilateral traders that wish to protect themselves against transmission price risk among two specific locations can do so in too ways. They can acquire transmission congestion contracts (TCCs) between the two locations. These are financial instruments underwritten by the ISO that entitles or (obligates) their holder to a payment that equals the nodal price difference between the nodes. Such a financial contract would enable a trader to fully hedge the transmission price risk between two nodes. Unfortunately, with prices being different at each node and the large number of different TCCs needed to enable full hedging for each possible bilateral transaction (a 3000 node system would require about 4.5 million of different TCCs) it is unlikely that a market for TCCs could achieve sufficient liquidity to make TCC pricing efficient. Without a liquid TCC market the value of these instruments as risk management tools for energy traders is questionable.

Bilateral traders can also manage transmission price risk by actively participating in congestion relief. The nodal pricing paradigm (as implemented for instance in the PJM pool) restricts such participation to incremental and decremental bids that can be readily interpreted within the framework of the central pool economic dispatch protocol. Specifically, a trader may submit incremental and decremental (inc/dec) bids that would allow the ISO to modify its injection as if it was a pool bidder. With demand side bidding it would also be possible to have inc/dec bids on the load side. Such inc/dec bids, however, allow the ISO to displace the bilateral generator by cheaper generation, for efficiency reasons, even if there is no congestion.

While proponents of nodal pricing based pools often argue that such systems (e.g. PJM or NYPP) can accommodate bilateral trading the reality is that only bilateral trading that fits within the rigid pool framework are allowed. It is not possible for a bilateral trader to limit its congestion risk without exposing its generators to "efficiency motivated displacement". A bilateral trader cannot submit an inc/dec bid and ask that it only be used if there is congestion. It is also not possible for a trader to cap its transmission cost by submitting a decremental bid or cap on the cost of transmission between two points (i.e. submit an inc/dec bid on the nodal price difference rather than on the nodal prices). Such a transaction cannot be decomposed into elementary sell/buy transactions within the pool and are, therefore, disallowed. Finally, as it was painfully realized when the PJM pool moved to nodal pricing, the popular user choice contracts, where the buyer can take delivery at any chosen location, cannot be protected against transmission price risk. Proponents of nodal pricing dismiss the disallowed bilateral transactions on the grounds that such transactions are socially inefficient and no one in his right mind would prefer them to the superior deals allowed by the pool. For instance: why wouldn't a bilateral generator submitting a decremental bid for congestion relief be willing to accept replacement of its generation by power priced below its declared decremental price? Such arguments expose the central planning mentality underlying the nodal price approach which does not tolerate bilateral trading that cannot be rationalized within that limited short-term efficiency perspective.

The complexity and restrictive characteristics of the nodal pricing framework are rationalized on the grounds of short-term economic efficiency. However, such efficiency claims hinge on unrealistic or simplistic assumptions. In particular, efficient resource use under a nodal pricing scheme requires that each generator and load bid their true costs or willingness to pay. This is a highly questionable premise given the inevitability of some locational market power in the electricity industry. Furthermore, intertemporal costs and constraints that are accounted for in unit commitment optimization affect optimal dispatch in a broad sense. The power flow optimization that is used to determine nodal prices is predicated on unit commitment decision. However, central unit commitment optimization is fundamentally incompatible with a competitive market (see Johnson, Oren and Svoboda [1997]). It requires information about costs and constraints that is either not revealed to the ISO or is subject to gaming (like in the UK system). None of the proposals or current implementations of nodal pricing offers a credible way of assuring optimal unit commitment. Indeed many deregulated electricity system (e.g. Norway, Victoria pool, California) opted for self-commitment where the decision to turn a unit on or off is left to the individual generators and not centrally optimized.

In drawing conclusions from the above discussion I like to point out an important distinction between precision and accuracy. While nodal pricing aims at providing precise locational economic signals their accuracy is questionable and hardly justifies the complexity and rigidity of that approach. The ability of this approach to meet its idealized efficiency objectives is hindered by unrealistic premises and by the fact that it ignores intertemporal considerations. Furthermore, the importance of accurate short-term economic signals toward achieving long-term efficiency is debatable. All this suggests that a less radical approaches that are suboptimal (second best) in terms of theoretical short-term efficiency but simpler and more transparent would be more desirable than the nodal pricing alternative.

Efficient Congestion Relief without Mandatory Economic Dispatch

Zonal pricing in conjunction with incremental and decremental (inc/dec) bids for congestion relief has been adopted in California as a less intrusive and simpler alternative to nodal pricing, which allows efficient congestion relief with minimal interference in the energy market. The basic principle of the California approach is to separate the energy market from congestion relief, which can be viewed as an ancillary service. Such separation empowers the ISO to use incremental and decremental bids by bilateral traders for the purpose of least cost congestion relief. The marginal congestion relief cost is then imposed on the serviced interzonal transactions as a congestion charge. The ISO, however, is not allowed to interfere in the energy market by using the inc/dec bids for the purpose of displacing "inefficient" generation beyond congestion relief needs.

The technical details of the California implementation are described by Papalexopoulos, Singh and Angelidis [1998]. I will illustrate here the basic principles of least cost congestion relief with a simple two-zone example. Figure 1 illustrates three interzonal bilateral transactions scheduled by three scheduling coordinators (SC). Each SC has its own supply and demand curve representing its load and generation capacity, which leads to its preferred schedule. In our example the preferred transaction quantity of each SC is chosen to maximize gains from trade (represented by shaded area). The corresponding cost of the marginal transaction unit varies across SCs (it is 17 for SC1, 15 for SC2 and 12 for SC3). It should also be noted that the energy settlement price of each SC depends on contractual arrangement and may differ from marginal cost. Each SC submits to the ISO a preferred schedule, which in this example is 1000MW. Since the interzonal transmission capacity is limited to 1500MW ISO intervention is needed to relieve congestion. In addition to a preferred schedule each SC submits inc/dec bids for congestion relief on both sides of the transmission line. The ISO's congestion relief protocol requires that the balanced schedule of each SC be preserved (an adjustment for losses is made but we assume no losses in our example). Hence the inc/dec bids of each SC on the two side of the congested interface can be added to provide a congestion relief bid or supply function. In our example we assume that the congestion relief supply function of each SC reflects the opportunity cost (or forgone trade gains) due to backing down the transaction.

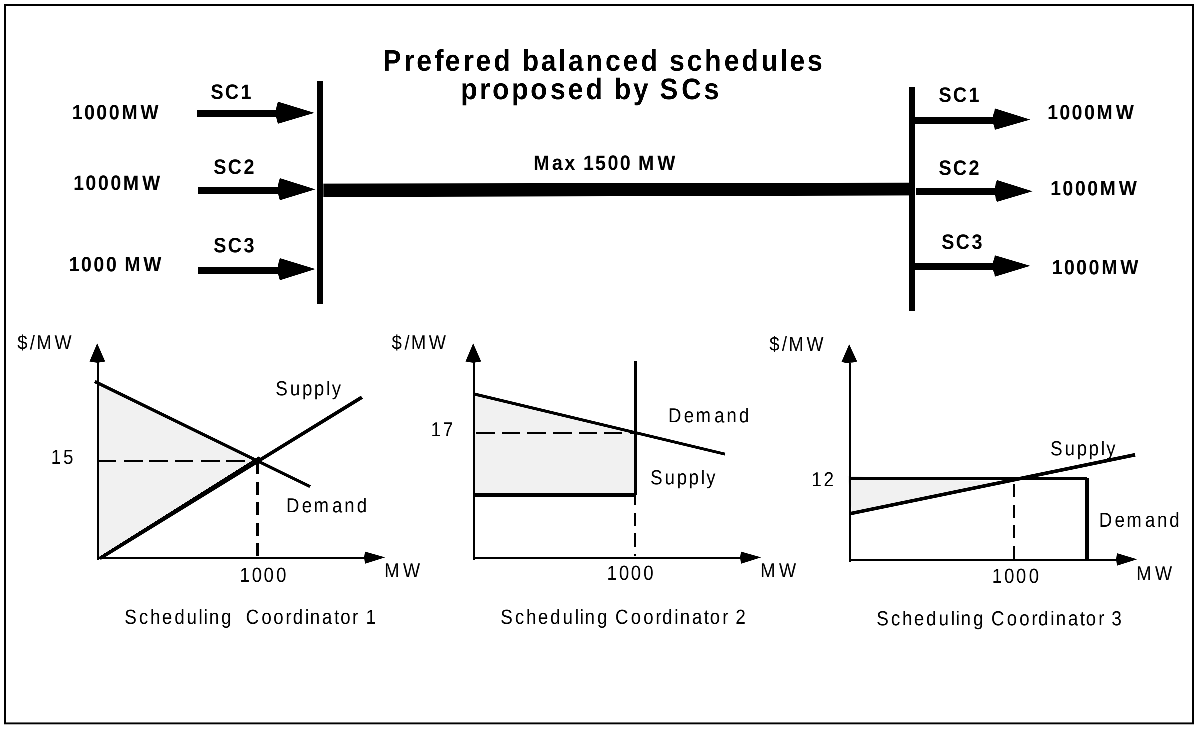

Figure 1

The Preferred Schedules of Three Scheduling Coordinators Exceed Transmission Capacity

Figure 2 illustrates how the congestion relief supply function are ideally determined by each scheduling coordinator and how they are aggregated by the ISO into an overall congestion relief supply function for the congested interface. Using the aggregate congestion relief supply function the ISO determines the marginal congestion relief price that will back off transactions across the interface to a total of 1500 MW (the interface limit). That marginal price is used to determine the curtailment level for each scheduling coordinator according to their individual congestion relief supply curves. The surviving transactions are charged a transmission congestion fee (per MWh) that equals the congestion relief marginal clearing price. One of the implications of the above approach is that we view congestion relief as a commodity or more precisely an ancillary service. It is also efficient and meets comparability criteria since the marginal congestion relief opportunity cost is equalized across all curtailed transactions. What this approach does not do is equalize marginal cost of generation at a node across scheduling coordinators (as economic dispatch would do). In other words, congestion is relieved with minimal intervention in the energy market.

Figure 3 shows the injections and withdrawals of the three scheduling coordinators after the congestion relief protocol has been enforced along with the supply and demand curve of each scheduling coordinator (those represent private information that is not disclosed to the ISO). We observe that SC2 who is still generating 800 MW has a marginal cost that is higher than some of the curtailed generation capacity of SC3. Note, however, that SC2 was lucky enough to find a buyer that is willing to pay more and hence had a higher opportunity cost for curtailment. Should they desire to do so, SC2 could make a deal with SC3 to buy up to 600MW and turn its own generation down to 200MW. Since SC3 can produce these 600MW cheaper than SC2 both scheduling coordinators could benefit from such a transaction. The adjusted injections after such a transaction are shown at the bottom of Figure 3. Note, however, that such trading among the scheduling coordinators is voluntary and since it does not involve using the transmission lines it does not involve the ISO. By contrast, in a nodal pricing approach, the ISO would have the authority and responsibility, to displace SC2's 600 MW with SC3 generation should the generators choose to participate in congestion relief by submitting decremental bids on the injection side.

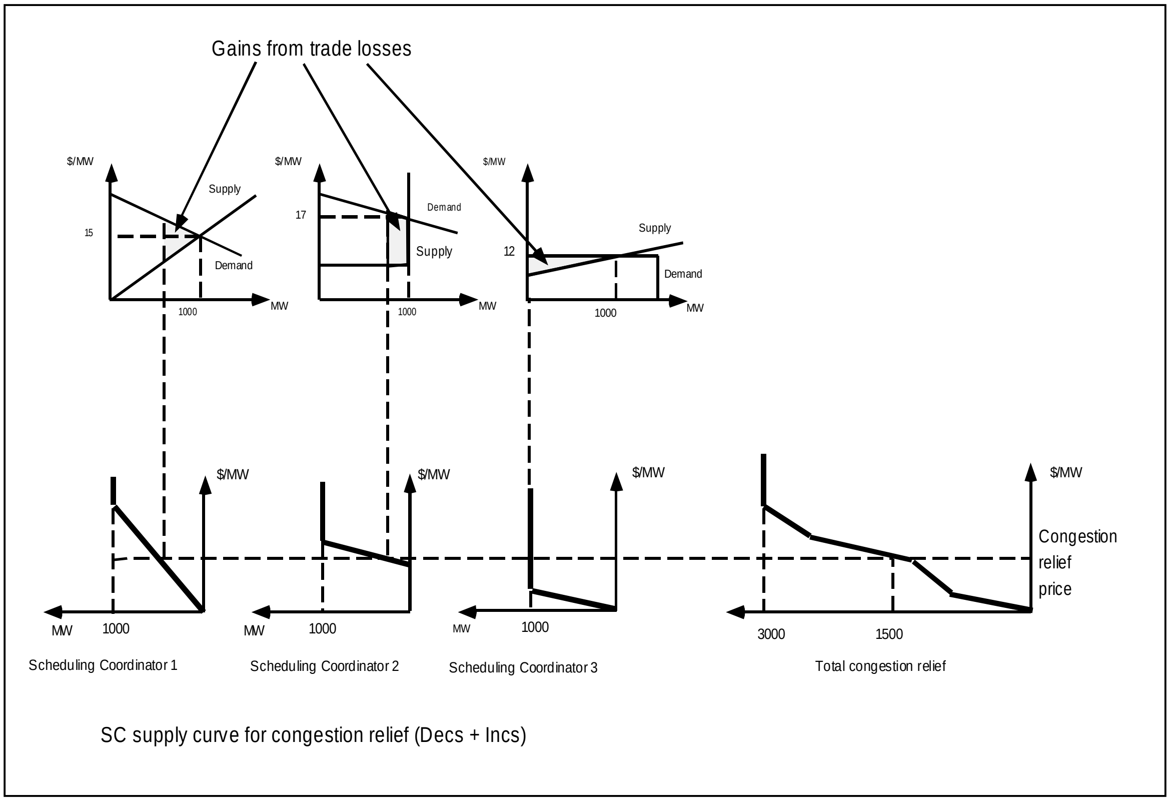

Figure 2

Scheduling Coordinators Submit Congestion Relief Bids Based on Their Opportunity Cost

Zonal Aggregation and Priority Access Pricing

Zonal aggregation has been adapted in many systems as a realistic alternative to nodal prices. The basic idea in this approach is to divide the grid into few congestion zones as illustrated in Figure 4, with separate spot markets whose respective market clearing prices set the uniform price within the zone. In California, for instance, two such zones have been designated for the purpose of pricing power on the demand side but on the supply side a finer resolution is used. When there is no congestion the zonal markets collapse into one. However, when congestion is present the zonal markets are decoupled and the zonal market clearing prices reflect the supply and demand conditions in each zone as well as the interzonal transmission capability. When interzonal congestion occurs, bilateral transactions across zones are subject to an ex-post congestion fee based on the congestion relief cost between the zone. Keeping the number of distinct zones small facilitates the formation of liquid markets for zonal futures and forwards which enable traders to hedge the interzonal congestion cost risk. In California, for instance, the price difference between COB and Palo Verde electricity futures reflects the basis risk for congestion between north and south. Thus trader can use these financial instruments to hedge interzonal transmission cost risk between northern and southern California. The protocol for interzonal congestion relief and determination of the congestion charge can be accomplished through the congestion relief bidding system described in the previous section.

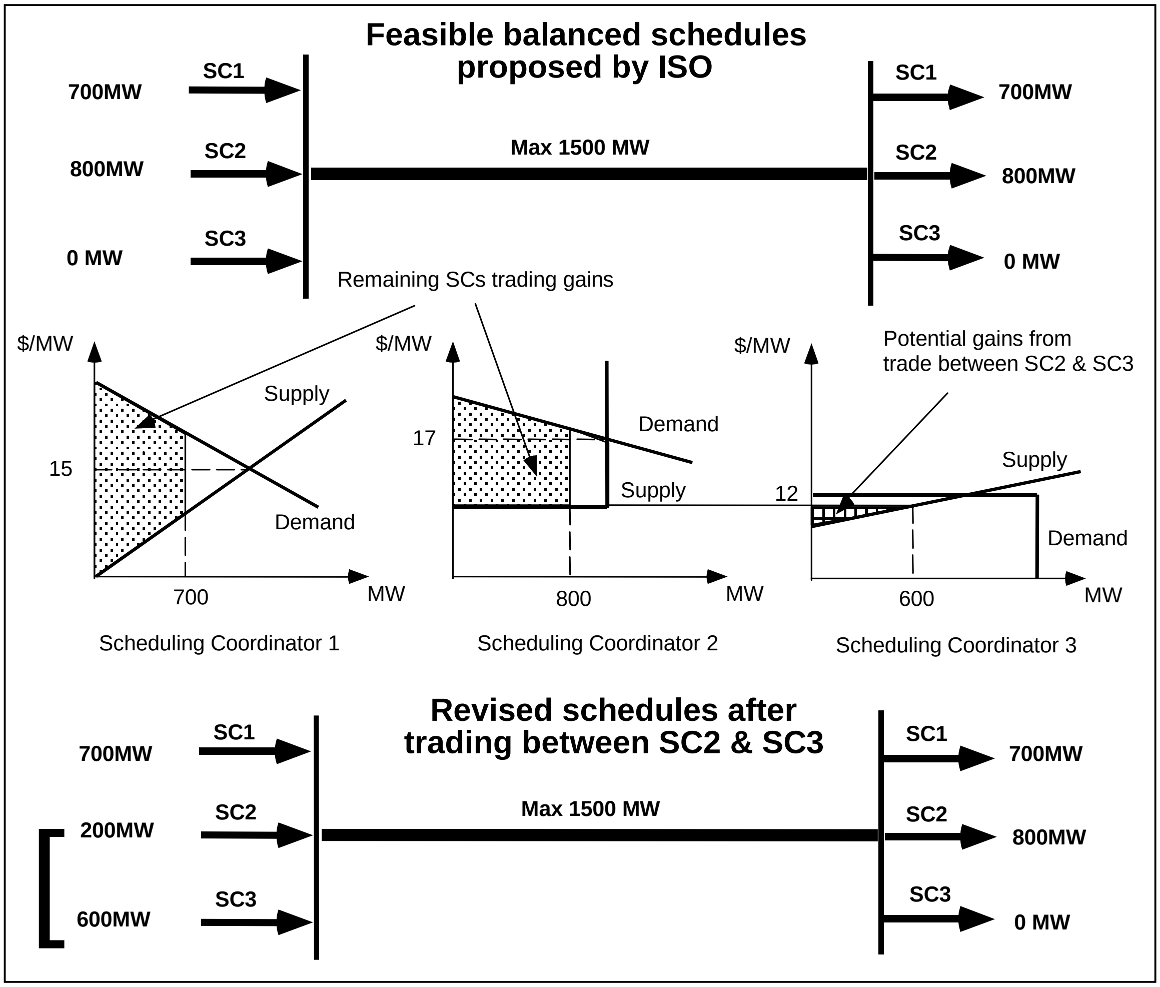

Figure 3

Least Cost Congestion Relief and Energy Trading among Scheduling Coordinators

The basic question is how many zones are needed? Hogan [1997] argues that node aggregation can only be justified when nodal prices are identical at the grouped nodes and in order to know when that condition is met one must compute nodal prices anyway. On the other hand, Green [1998] shows using stylized data from the UK that the efficiency losses due to zonal aggregation are relatively small. The main issues that must be addresed when the number of zones is small (relative to the size and complexity of the network), are intrazonal congestion management and the provision of locational economic signals within the zone. In recent work Deng and Oren [1998] propose a new approach for addressing these issues by offering a uniform ex-ante priority differentiated zonal network access tariff in conjunction with an ex-post interzonal congestion fee. The proposed zonal tariff works as a zonal postage stamp method based on the injection zone, i.e., it is uniform regardless of the specific injection node within the zone. However, traders have the option to select among a variety of "stamps" which will determine their priority (place in line) should congestion occur (e.g. first class stamp, second class etc). While the stamp prices are uniform within the entire zone the actual probability of service they entail will depend on the location of the transaction. Consequently, traders can choose how much they wish to pay for transmission access by trading off the cost of the service priority against the locational risk of curtailment and their opportunity cost associated with curtailment. It is reasonable to expect that a high valued transaction or a transaction that impacts congestion prone links will opt for a higher priority level to reduce curtailment risk. On the other hand a transaction that is unlikely to be curtailed due to its location will opt for the least cost service. The ability of traders to self-select their service priority levels results in correct economic signal (direction wise) for efficient rationing when transmission capacity is scarce. It also produces the correct signal for location of new generation assets, since locating such assets where they do not impact congestion will allow their owners to save on transmission cost by selecting a lower service priority.

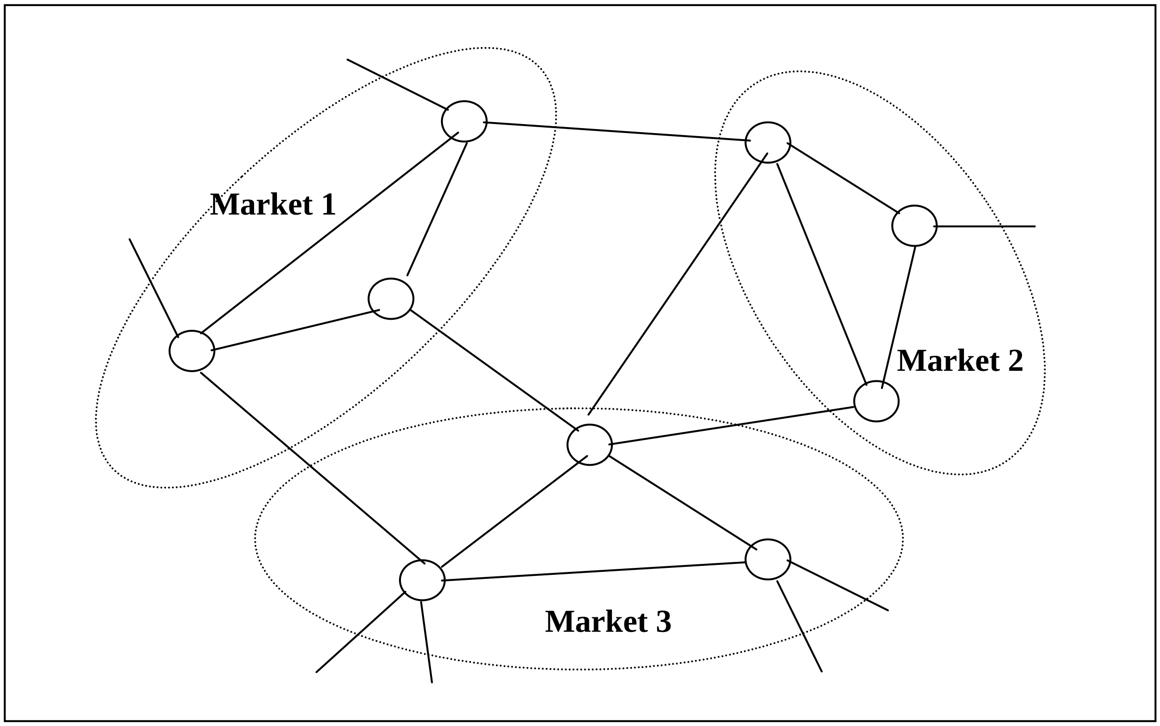

Figure 4

Zonal Aggregation of Nodes

There are many possible ways of implementing a priority network access tariff. The Pricing scheme proposed by Deng and Oren [1998] takes the form of an "option insurance". The basic premise of the proposed scheme is that the spot prices in each zone serve as a reference for replacement of curtailed delivery into the zone or for financial settlement for unfulfilled delivery contracts. So for example, if a trader holding a contract for delivery of power from zone 1 to zone 2 is curtailed his cost per curtailed MWh is the settlement or replacement cost in zone 2 (S_2) less his saved marginal cost of generation (v). In addition, we assume that an interzonal transaction is subject to an ex-post congestion fee per MWh that equals the difference of zonal spot prices (S_2-S_1). That fee is avoided when the transaction is curtailed. Hence, the trader incurs a net curtailment cost per MWh of (S_2-v) - (S_2-S_1)= S_1-v. It follows that the opportunity cost of a curtailed transaction amounts to the opportunity cost of selling the power into the zonal spot market where the generator is located. Since the generator has the option not to generate when the spot market price is below his marginal cost, the opportunity cost of a generator in zone 1 of not having physical access to the network is Min[0, S_1-v]. This opportunity cost is the same as the payoff of a call option with strike price v, with respect to the zonal spot price. Given the uncertainty in spot prices, the forward value of network access to a generator in zone 1 with marginal cost v is given by the actuarial value of a call option with strike price v. Figure 5 illustrates the fact that the value to a trader of transmission access is the same whether the transaction is interzonal or intrazonal due to the ex-post congestion fee imposed on interzonal transactions.

Based on the above observation we can construct a zonal transmission access tariff that takes the form of an insurance premium for insuring the option to sell power in the zonal spot market. The access fee collected by the ISO, is set to the actuarial value of the option , X(v)=E{Min[0,S_1-v]}which depends on the insurance level that is defined by the strike price v (X(v) is a decreasing function of v). Purchasing insurance level v for one unit of injection entitles the trader to either physical access to the local grid or to compensation in the amount of the revealed opportunity cost, i.e., Max[0,S_1-v]. The ISO can then manage congestion so as to minimize compensation payments to curtailed transactions, net of the interzonal congestion fee revenues. Thus a higher insurance level (lower v) entails a higher priority, which at the same location implies a higher probability of network access. The access probability corresponding to the same insurance level varies, however, across locations and over time depending on the network load conditions.

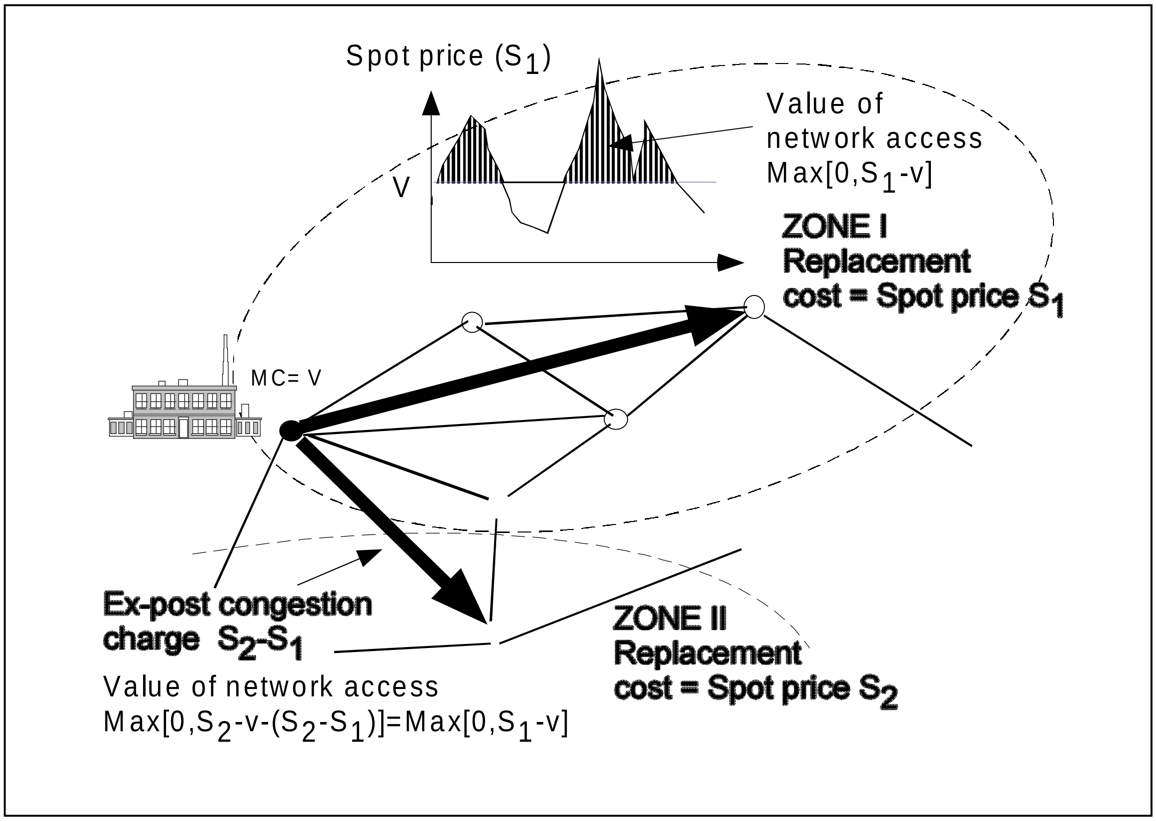

Figure 5

Valuation of Network Access as An Option

It can be shown that under the above scheme, if every trader insures its transactions revealing the true marginal cost, then minimizing net compensation is equivalent to economic dispatch. However, traders will tend to balance the risk of curtailment with the gain from a lower transmission cost and will have incentives to share the risk of curtailment and underinsure (select a strike price higher than the true marginal cost) in order to lower transmission access payments. This incentive causes distortion of the economic signal. Such distortion is the price we pay for having a uniform tariff across the entire zone (differentiated only according to priority) rather than node specific pricing. Nevertheless, the economic signals are in the right direction and preliminary simulations described by Deng and Oren [1998] suggest that the efficiency losses due to "inaccuracy" of the economic signals are modest.

Conclusion

Short term theoretical efficiency claims of the nodal pricing approach to transmission tariffs and congestion management are based on unrealistic assumptions and a myopic view of optimal resource use. Hence, such claims do not justify the burden of thousands of different prices, the drawbacks of ex-post transmission charges and the constraints imposed on bilateral transactions and on risk management. Zonal aggregation and decoupling of the energy market from congestion relief protocols simplify transmission pricing and congestion management without substantial short-term efficiency losses. Such simplifications offer more certainty of transmission cost to traders and better risk management capability. They also enable a higher degree of decentralization in the energy market, which will promote innovation and long-term efficiency.

References

Deng Shijie and Shmuel S. Oren, "Priority Network Access Pricing for Electric Power", POWER Conference, University California Energy Institute, Berkeley, California (March 1998).

Green Richard, "The Economic Impact of Transmision Pricing Schemes," Working paper presented at INFORMS meeting, Montreal Canada, (April 1998)

Johnson Raymond B., Shmuel S. Oren and Alva J. Svoboda, “Equity and Efficiency of Unit Commitment in Competitive Electricity Markets,” *Utilities Policy*, Vol. 6, No 1, (1997) pp. 9-19.

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

Hogan William W., "Nodes and Zones in Electricity Markets: Seeking Simplified Congestion Pricing", 18th Annual North American Conference of the USAEE/IAEE, San Francisco California (September 1997).

Papalexopoulos Alex, Harry Singh and George Angelidis, "Congestion Management by an Independent System Operator," POWER Conference, University California Energy Institute, Berkeley, California (March 1998).