

**Competitive Electricity Markets:  
Why They Are Working  
And  
How To Improve Them**

**Larry E. Ruff**

**n/e/r/a**

**12 May 1999**

# Competitive Electricity Markets

## Table of Contents

<b>1. INTRODUCTION .....</b>	<b>1</b>
1.1 THE BASIC MESSAGE.....	1
1.2 THE BASIC LOGIC.....	2
1.3 THE CURRENT SITUATION WORLDWIDE .....	3
1.4 THE DIRECTION OF CHANGE .....	4
<b>2. MARKET PRINCIPLES APPLIED TO ELECTRICITY .....</b>	<b>6</b>
2.1 REQUIREMENTS FOR MARKET EFFICIENCY .....	6
2.1.1 A Supportive Political/Legal/Social Framework .....	6
2.1.2 An Infrastructure of Monopoly or Non-Market Facilities .....	6
2.1.3 Well-Defined and Enforceable Property Rights .....	6
2.1.4 Practical Ways To Trade Rights and/or Price Interactions .....	7
2.2 THE FUNDAMENTAL TRADE-OFFS IN MARKET DESIGN.....	7
2.3 WHY COMPETITION CAME LATE TO ELECTRICITY .....	8
2.4 THE BREAKTHROUGH: ISOs, SPOT MARKETS AND FINANCIAL CONTRACTS .....	9
<b>3. ALTERNATIVES FOR CONTRACTING AND REAL-TIME TRADING.....</b>	<b>12</b>
3.1 “PHYSICAL” VERSUS “FINANCIAL” CONTRACTS .....	12
3.1.1 Ownership or Title.....	12
3.1.2 Commercial Terms.....	13
3.1.3 Insulation from the Spot Price, or Trading Outside the Pool.....	13
3.1.4 Contracts for Non-Energy Physical Services or “Green” Energy .....	14
3.2 THE CRITICAL ISSUE: SHORT-TERM TRADING ARRANGEMENTS .....	15
3.2.1 Contract Damages and Remedies.....	15
3.2.2 Expanding – and Limiting – Market Choices .....	16
3.3 CENTRALIZED VS. DECENTRALIZED SPOT TRADING IN ELECTRICITY.....	17
3.3.1 The Short-Term Coordination and Trading Options .....	18
3.3.2 Option 1: Separated Short-Term Trading and Non-Market ISO Processes .....	19
3.3.3 Option 2: Financial Contracts and a Centralized Spot Market .....	20
3.3.4 The Natural Monopoly in Real-Time Market-Making.....	21
<b>4. SUCCESSES AND PROBLEMS IN ELECTRICITY MARKETS.....</b>	<b>23</b>
4.1 MANAGING TRANSMISSION CONGESTION.....	23
4.1.1 The Problem: Externalities Caused by Grid Congestion .....	23
4.1.2 The Logical Solution: Congestion Pricing .....	25
4.1.3 The Trend: Locational Energy Pricing .....	26
4.2 INVESTMENT IN GENERATION CAPACITY.....	27
4.2.1 The Problem: “Free Rider” Network Externalities.....	28
4.2.2 Logical Solution 1: Installed Capacity Requirements/Prices .....	30
4.2.3 Logical Solution 2: An Hourly Energy Price “Adder” .....	31
4.2.4 The Trend: No Clear Direction but Increased Recognition of the Problem.....	33
4.3 RETAIL COMPETITION FOR SMALL CONSUMERS .....	34
4.3.1 The Problems: High Costs and Limited Competition for Small Consumers.....	35
4.3.2 The Logical Solution, Part 1: Correctly Define the Objectives .....	35
4.3.3 The Logical Solution, Part 2: Spot Prices and Financial Contracts.....	37
4.3.4 The Trend: Slow Progress with Mixed Results .....	38
<b>5. CONCLUSIONS.....</b>	<b>40</b>

# **Competitive Electricity Markets: Why They Are Working and How To Improve Them**

**Larry E. Ruff**

**n/e/r/a**

**12 May 1999**

## **1. INTRODUCTION**

The competitive revolution in electricity, while not without its problems, has had real benefits wherever it has been tried. But there is no reason to be complacent about the future of the competitive revolution. Counterrevolutionary forces are already on the move, stimulated as much by the successes of the revolution as by its shortcomings. This paper presents an explanation of both the successes and the shortcomings of competition in electricity, with the objective of helping assure that the future is one of productive evolution rather than unproductive reaction.

### **1.1 THE BASIC MESSAGE**

The basic theme of this paper is that the rather surprising success of competition in electricity has been due largely to the development of spot markets that are integrated with real-time physical dispatch. Such an integrated spot market/dispatch process is the only practical way to price or internalize the real-time network externalities that otherwise make competitive electricity markets unacceptably inefficient and unreliable. The financial contracting that becomes possible only when there is an open spot market then largely displaces more complex physical contracting, allowing producers and consumers to meet their commercial needs with relatively low transaction costs and risks.

Most of the problems that have arisen in electricity markets – other than those due to structural problems such as inadequate competition – are attributable to specific flaws in the integrated spot market/dispatch process or to failure to take full advantage of the spot prices arising from this process. In particular, problems related to transmission congestion and peaking capacity arise when system prices do not adequately internalize network effects. And the cost of retail competition for small consumers is greatly increased by failure to give such consumers access to spot market prices so that they, too, can get the benefit of financial contracting. If competition in electricity is to reach its full potential to provide benefits to consumers, it will be necessary to implement more sophisticated spot markets that better reflect network complexities and to give all consumers, even small ones, access to spot prices and financial contracts.

Unfortunately, the features that make integrated spot market/dispatch processes so successful – low transactions cost and market transparency and efficiency – are the very features that are provoking the counterrevolution. A transparent, efficient spot market allows producers and consumers to deal directly with each other with less need for middlemen and market makers, and

helps new, small, niche players compete effectively with established, large, diversified players. Not everybody finds these features to their liking or advantage.

So there is no shortage of powerful critics attacking the integrated spot market/dispatch process/financial contract model. These critics continue to delay introduction of integrated spot market/dispatch processes in most of the United States and Europe, persuaded California to try to force an inefficient split between the dispatch process and spot market(s), and are trying to persuade the British virtually to scrap the Pool that, for all practical purposes, started it all.<sup>1</sup>

Despite the reactions it has provoked, the basic model of an integrated spot market/dispatch process with financial contracting is so logical and is proving so successful that – barring some unexpected technological breakthrough that fundamentally changes the nature of electricity systems – it will continue to be the best or even the only way to create effective and efficient competition in electricity. Many variations on this basic model are being and will continue to be tried, and all of them will continue evolving as new problems arise. None of them are or ever will be perfect or simple, because a real electricity system is and always will be imperfect and complex. But nobody has proposed another practical way to internalize the complex externalities that inevitably arise during real-time operations of an integrated electricity system. And nobody has demonstrated or even plausibly argued that competition can be effective and efficient if large network externalities are left unpriced and unmanaged.

## **1.2 THE BASIC LOGIC**

Technical logic dictates that an electricity system have a central process to co-ordinate real-time physical operations; to the extent that this process is not based on markets, it must be based on less efficient command-and-control methods. Economic and commercial logic requires that a commodity market have short-term trading arrangements to bring market positions into agreement with physical reality; to the extent that market trading does not reflect physical reality, some non-market process must close the gap between the market and reality. These two propositions imply that the best way to maximize the role of the market and minimize the role of non-market processes is to base real-time physical operations on a spot market and to allow market participants to use this market for commercial purposes to the extent they find this useful.

If market participants have the option of trading in a spot market that reflects physical reality, they will naturally tend to trade under contracts that specify monetary payments based on the spot prices rather than under contracts that specify physical actions. Any two parties who want to trade under physical contracts can do so, but will be at a disadvantage relative to competitors using the more flexible and efficient financial contracts and hence will quickly learn better. Financial contracting can even reduce the high transaction costs of supplying small consumers if these consumers have access to the spot price.

---

<sup>1</sup> Norway, Chile and “tight” US power pools had generator-only markets years before the England and Wales Pool began operating. But the England and Wales Pool was the first that was open to buyers as well as sellers. It is such an open, two-sided market that makes real competition possible.

The alternative to integrating spot trading with physical operations is to force a separation between trading and operations. Multiple, competing markets divorced from physical operations can determine prices and quantities that ignore transmission constraints and real-time events, and then the system operator – necessarily a monopoly – can use non-market methods to close the gap between the hypothetical market solution and physical reality. But if network externalities are large and conditions can change quickly, multiple markets divorced from physical operations will be unable to price important network interactions and last-minute events accurately, and the gap remaining for the system operator to close will be large. If the system operator closes this gap in the most logical and efficient way – i.e., with a market – market participants will start using this market for commercial trading and the separation between the market and physical operations will be lost. The only way to maintain such a separation is to require the system operator to use only inefficient, non-market processes to manage network externalities and last-minute events.

Although it may seem perverse to require the system operator to use inefficient, non-market processes to solve what is essentially an economic problem, there are many advocates of such an approach. For example, the regulator (OFFER) and government in England and Wales are proposing to eliminate the Pool; California has forced an artificial split between the Independent System Operator and the Power Exchange; and there is strong opposition in much of the US and Europe to the idea that the system operator should use open spot markets to manage real-time operations. Handcuffing the system operator in this way creates a lot of business for private marketers and gives strong competitive advantages to the largest, most diversified market participants. But such non-market operations cannot support efficient or effective competition on any complex system with more than a few players.

Both logic and history suggest that improved competition in electricity will require more, not less, sophisticated spot market/dispatch processes that reflect more, not less, of reality, so that the resulting prices internalize more, not less, of the network externalities. This is the best way to make the necessarily monopoly system operator more of a mechanic tending the market-clearing machinery and less of an active intervenor in the market, so that the results of competition will depend more on market forces and less on decisions of the monopoly central operator. As both a logical and a historical matter, forcing markets to be simple when reality is complex is a big step in the wrong direction.

### **1.3 THE CURRENT SITUATION WORLDWIDE**

On the whole, first-generation electricity markets have worked surprisingly well, especially considering that when they were designed most electricity experts thought a market-driven electricity system was impossible and there was little theory to light the way. Although many small and a few large mistakes were made in the design of these first-generation markets, the trading and operating arrangements have generally been adequate to allow state monopolies to be broken up and (in most cases) privatized, resulting in early and obvious improvements in efficiency within individual enterprises. But problems have arisen that require changes in trading and operating rules.

Some electricity markets that are working reasonably well are in danger of becoming victims of their own success. In England and Wales, for example, competition has reduced costs and improved performance, but has put severe competitive pressure on the coal industry, while the Pool trading arrangements are efficient enough to make it hard for middlemen and market makers to earn high transaction fees. The politically potent combination of Yorkshiremen brandishing their picks and Citymen wielding their computers may result in significant, and not necessarily positive, changes in the Pool trading arrangements.

In some cases, the apparent initial success of competition in electricity has been due to the existence of excess transmission and generation capacity,<sup>2</sup> newly created entities who could be pressured into behaving themselves, and a compliant base of (usually small) consumers who could be taxed to pay the transition costs. Over time, however, inherited capacity depreciates while demand grows, more ruthless competitors get into the game, and there is pressure to let even small consumers share in the benefits of competition. The problems resulting from weaknesses in the first generation markets become more serious, and the trading and operating arrangements must be modified.

When an electricity market discovers a problem that can no longer be ignored, the first reaction is usually to direct the monopoly system operator to solve it somehow. For example, the England and Wales Pool gave the monopoly system operator profit incentives to manage transmission congestion and within-day events outside the market. NEPOOL (the New England Power Pool) allowed the system operator to forbid new generator connections where they added to congestion problems. Victoria (Australia) directed the system operator to contract for peaking capacity and spread the cost across all users. But such non-market solutions usually create even more problems than they solve, creating pressure to develop a more market-based solution.

Second generation markets have generally tried to create more sophisticated market arrangements that better reflect and price reality so that the role of the system operator is reduced. For example, the markets in New Zealand, Argentina, PJM (the Pennsylvania, New Jersey, Maryland Power Pool), the New York Power Pool, Australia and California are more sophisticated than the markets designed earlier. Most notably, these later markets give participants more latitude to determine their own operations, use real-time information to determine prices (which means prices are not known until some time after the fact or "*ex post*"), and include in prices the effects of at least some transmission constraints. All of these advances require a spot market closely integrated with physical dispatch; they are virtually impossible in external markets separate from dispatch.

#### 1.4 THE DIRECTION OF CHANGE

As electricity systems and competitive markets mature, the trading arrangements become more sophisticated in two, seemingly contradictory, directions. On the one hand, the system operator

---

<sup>2</sup> A notable exception is in Latin America, where electricity systems were privatized and made competitive primarily to stimulate badly needed investment. These objectives have largely been accomplished in most cases, albeit sometimes at the cost of stimulating too much investment.

imposes fewer constraints on generators (and price-responsive loads, which in this paper are implicitly included in most references to “generators”) so that market participants themselves can make more of their own operational decisions and then live with the results. On the other hand, the system operator administers more sophisticated market mechanisms that are more closely integrated with real-time physical operations.

It is no contradiction to say that the system operator should give market participants more freedom to make their own decisions and at the same time should administer more sophisticated real-time markets. Indeed, a logical requirement for giving market participants more freedom to respond to prices is that the prices to which they respond must better reflect physical realities, including the complex realities of the transmission system and real-time events. Letting market participants do whatever they want in response to prices that do not reflect real-time physical reality would be illogical, inefficient and ultimately impossible.

If the real-time prices faced by generators reflect actual real-time network effects, each generator can be allowed wide latitude to decide for itself when to start or stop its units based on its own forecasts of real-time prices and on its contracts.<sup>3</sup> System prices can then be determined based on what actually happens in real time. Such a market simplifies and minimizes the role of the system operator, who does not project, optimize or assume risks over multiple hours. Each market participant decides for itself how it wants to contract and operate in response to its own projections of real-time prices and its own intertemporal constraints such as unit start-up times and ramping rates. The system operator is left with the relatively simple task of clearing the short-term spot market and managing events within each market period, given the generating resources that are making offers and the uncontrollable demand that must be met.

If desired, the system operator can also operate one or more forward markets, say a day ahead and then an hour ahead of real time, to help market participants predict and hedge against real-time prices. But even if the system operator does not provide this service, private marketers will develop contracts with terms ranging from years to hours ahead that can be traded in various decentralized, *ex ante* markets, and market participants can use these instruments to hedge against uncertainty in the real-time prices. Marketers, arbitrageurs and speculators will assure that the prices in the various *ex ante* decentralized markets and the central market(s) operated by the system operator converge to economically consistent values. Such decentralized market activity should be encouraged – but not by forcing the real-time market to be inefficient just to guarantee a large and profitable role for middlemen.

---

<sup>3</sup> It is not usually possible to allow large generators to operate as they please without informing or following instructions from the system operator. But generators can be allowed to submit simple information and spot market offers/bids to the system operator every few hours or less, which can be used to determine dispatch schedules more-or-less continuously. Generators will learn to submit information and bids/offers that result in the dispatch instructions they want. As long as real-time system prices reflect real system costs and benefits, such anticipatory bid/offer behaviour is desirable and efficient.

## **2. MARKET PRINCIPLES APPLIED TO ELECTRICITY**

Markets are wonderful things for allocating scarce and hence valuable resources and commodities. Adam Smith was surely right that the invisible hand of supply and demand guides the uncoordinated army of bakers to deliver the right amount and type of bread at the right time and place without any conscious, central control. Smith's reasoning can be extended to demonstrate that the invisible hand can work even in electricity, guiding electricity producers and consumers to do (approximately) the right things. But markets work well only when property rights are well-defined and can be traded and priced efficiently – requirements that are not always met without help from some very visible hands. Real competition in electricity is possible only to the extent that conscious efforts are made to develop property rights and markets that can deal adequately with the strong and nearly-instantaneous network interactions that characterize an integrated electricity system.

### **2.1 REQUIREMENTS FOR MARKET EFFICIENCY**

Markets do not themselves design, implement, or pay for everything they need to operate successfully. Particularly when new resources are being brought into the market and externalities are strong and complex, non-market processes play a large role in defining the role and form of markets and in providing much of what markets need to function. It is worth reviewing some of the areas in which markets need assistance from non-market processes.

#### **2.1.1 A Supportive Political/Legal/Social Framework**

Markets do not work well if the political environment is unstable or corrupt, if property rights are poorly defined or difficult to enforce, or if society in general is hostile to markets – although when these conditions are not met, no alternatives to markets work very well either. A supportive political, legal and social framework is generally taken for granted or is regarded as implicit in the other conditions listed below. But it is worth remembering that markets do not work well if some very basic conditions are not met

#### **2.1.2 An Infrastructure of Monopoly or Non-Market Facilities**

Markets cannot function without the support of a legal, social and physical infrastructure that is not and largely could not be produced by markets themselves. The legal system, transport facilities, educational institutions, etc., were all created by historical social processes that are outside the market and are sustained by the very visible hand of governments. Market processes can and should be used to help provide and maintain this infrastructure efficiently, but without the guidance and support of non-market social processes will do so only partially and inefficiently. Markets simply do not create for themselves all the infrastructure they need to function well.

#### **2.1.3 Well-Defined and Enforceable Property Rights**

Markets do not function well without well-defined and enforceable rights to collect the benefits, and to be compensated for the costs – i.e., to internalize the benefits and costs – created by

producing and trading scarce resources and costly services. Economic history is largely the story of what happens when some previously non-scarce and hence non-owned or “common” resource – wild animals and plants, the village pasture, mineral deposits – becomes scarce and hence valuable, or when the previous owner/controller of a resource is dethroned or otherwise loses control. It can take years of conflict to define what the property rights are, decide who owns them, and develop workable markets in which they can be traded. The social process of defining and allocating property rights is not finished yet even for traditional commodities, much less for clean air, flowing water, satellite orbits, computer software, internet information – or energy and transmission on an integrated electricity grid.

As electricity markets continue to develop, it will be necessary to define, allocate, price and trade ever-more-sophisticated property rights. These rights will not define themselves or emerge in socially acceptable form from private negotiations among the most directly interested parties. People who understand the needs of traders, the technical realities of the integrated electricity system and the public interest in efficient and open markets will have to come together in organized processes to decide what kinds of property rights make sense. These processes will often have to be conducted under the auspices of government, if only because legislatures and courts will have to resolve disputes over who initially owns newly defined property rights. And everybody operating on the same interconnected grid will have to use the same basic definition of property rights.

#### **2.1.4 Practical Ways To Trade Rights and/or Price Interactions**

Markets can allocate resources efficiently only if rights to property and its fruits can be traded freely and at low cost, so that market-clearing prices and quantities can reflect all significant costs and benefits resulting from physical actions. For most common commodities, markets can develop freely and in competition with each other, so that the “market in market making” can determine the best form and scope of markets. But when externalities are strong and complex and the markets must clear very quickly, the natural end result of competition among markets may be a single market. In such cases the market process itself is a natural monopoly that requires some degree of social design and regulation to assure that all traders have equal and non-discriminatory access to the market and that the interests of affected parties not directly acting in the market – particularly small consumers –are protected.

## **2.2 THE FUNDAMENTAL TRADE-OFFS IN MARKET DESIGN**

The conditions cited above under which markets work well are clearly not absolute, and will always be met only more-or-less well or poorly. But the obvious fact that markets are always imperfect does not demonstrate that there is a better alternative, much less define what such alternatives might be. Conversely, just because it is possible to create a market in something does not prove that it is worth the costs of doing so, rather than doing something else or nothing. Deciding when and how to institute a market involves complex trade-offs.

As an economy grows and becomes more complex, previously unmanaged common resources become more scarce and valuable, and interactions among market participants become more serious and must be managed somehow. In most real-world situations, some evolving

combination of the following, progressively more sophisticated, management techniques is used to manage interactions.

- (a) **Toleration of Unmanaged Effects:** When externalities first appear, the inefficiency and unfairness resulting from unpriced and unmanaged effects are simply tolerated; for example, air pollution for a long time was, and often still is, left unmanaged within wide limits.
- (b) **Non-Market Management by a Monopoly:** When the externalities become too important to ignore, a government agency or (presumably regulated) private monopoly may be charged with managing interactions by non-market means; for example, pollution authorities set emission limits for specific sources based on technical criteria, with little help from markets.
- (c) **Creation of Property Rights, Prices and Markets:** When the inefficiencies resulting from non-market management become too great, tradable property rights and/or pricing mechanisms are developed to capture some of the otherwise unpriced costs and benefits; for example, tradable emission limits and emission taxes are becoming more common in pollution management.

All real-world markets involve some continuously evolving combination of these three options. But these are the only options. It is not logical or helpful to insist simultaneously that something must be done about some complex externality, that the solution must be a market rather than monopoly management, and that the market must be simple. If reality is complex, that complexity will have to be dealt with somewhere – if not explicitly in the market, than hidden within some monopoly or simply tolerated as inefficiency. There are no other choices.

### 2.3 WHY COMPETITION CAME LATE TO ELECTRICITY

The trade-offs among tolerating externalities, letting a monopoly manage them, and creating efficient rights and prices to internalize them are more severe for electricity than they are in more traditional markets that have developed “naturally”. Because all market participants are connected to the same grid on which power flows according to complex physical laws, externalities are potentially very strong and prices that fully and accurately internalized them all would be impossibly complex. The difficulty of pricing interactions on the common grid, more than the economies of scale in generation, prevented the development of competition in electricity until recently, and still limits the extent to which markets can replace monopoly control.

The first electricity systems were small, isolated monopolies with a few generating units under central control. Gradually these isolated systems were interconnected to allow what an economist would call trade, but what the engineers thought of as reserve sharing. The economists’ jargon “externalities” had not been invented yet, but the engineers knew that the uncoordinated operation of any significant connected entity could create catastrophic costs for all of them. As interconnections grew, so did the scope and complexity of the monopoly control system. Nobody knew how to price real-time grid interactions, but there was little reason to do so because neighboring monopolies essentially bartered reserves and energy in kind and cooperated to control the operations of all connected facilities to maintain reliability.

Eventually, simple contract trading developed among neighboring monopolies, but still without any actual pricing of network interactions. The selling monopolist would agree to deliver a certain amount of energy to the buying monopolist for an agreed price when asked to do so, and the buyer would simply integrate this source into its unpriced central dispatch process. As pressure for competitive generation developed in the United States in the 1970s and 1980s, this type of contract was extended to independent power producers (IPPs), so that an IPP could be paid for delivering energy to the local monopoly utility. But real competition, in which a generator could compete to sell directly to a consumer or to a distributor/retailer not affiliated with the local monopoly utility, did not develop until the 1990s.

For at least the last half-century, the principal obstacle to real competition in electricity has not been scale economies in generation, but the absence of market-clearing system prices to internalize complex network externalities. As long as only a few, cooperating monopolies are using the system, the externalities can be crudely controlled by simple administrative rules without much concern about the level or allocation of the resulting costs and benefits. But as competitors try to get into the market, the level and particularly the allocation of costs and benefits become critical. Without markets to internalize complex network externalities some regulated monopoly must manage them, and competition is stalled before it really begins.

This process has been playing itself out in the United States over the 1990s. Early in the decade, the Federal Energy Regulatory Commission (FERC) required monopoly utilities to provide “open access” to their grids, but did not define what this meant. When nothing happened, FERC required integrated monopoly utilities to post open-access tariffs specifying regulated prices for various services such as imbalance energy and ancillary services, but did not define how such prices were to be determined. The resulting complex tariffs have allowed some contract trading among large entities, but with increasing inefficiencies and conflict as more competitors try to get into the game. The industry and FERC are now embroiled in the ultimately hopeless task of creating administrative rules to allocate scarce resources in the absence of well-defined property rights and real market-clearing mechanisms.<sup>4</sup> Relatively effective and efficient competition is emerging from this chaos only where the FERC-required *pro forma* open-access tariff has taken the form of a spot market operated by an independent system operator.

## 2.4 THE BREAKTHROUGH: ISOS, SPOT MARKETS AND FINANCIAL CONTRACTS

The breakthrough that made real competition possible on an electricity grid was the concept of an independent system operator (ISO) that operates a centralized spot market more-or-less integrated with real-time physical operations and open to competitive buyers, while managing all

---

<sup>4</sup> For example, the non-market “transmission loading relief” (TLR) procedures that are used to curtail some bilateral transactions when transmission is congested are strongly – and accurately – criticized as unfair, inefficient, arbitrary, discriminatory, etc. Letting traders “buy through” such administrative curtailments will help, but in the absence of well-defined rights and efficient markets will be at best a partial and inefficient solution.

significant unpriced effects as a non-discriminating monopoly.<sup>5</sup> The England and Wales Pool (1990) was the first well-known and widely imitated such system, but many others have been or are being developed around the world.

During the 1990s, electricity markets based on an ISO and an open, centralized spot market have been established in Norway (since expanded to include most of Scandinavia), Argentina, New Zealand, Colombia, Peru, Ukraine, Victoria (and subsequently all of eastern mainland Australia), Spain, Alberta, California, PJM (the Pennsylvania, New Jersey, Maryland power pool) and elsewhere. Similar markets will soon be operating in New York, New England and Ontario, and will eventually emerge elsewhere.

The principal function of an ISO-administered integrated spot market/dispatch process is to price – i.e., to internalize – network interactions and real-time events more accurately than is possible in decentralized markets estranged from physical reality. Such pricing or cost internalization by the ISO in no way reduces the commercial importance of bilateral contracts freely negotiated between buyers and sellers. Indeed, one of the principal advantages of an ISO-administered spot market is that it greatly simplifies bilateral contracting by providing reference prices that reflect physical reality. Once the spot market is there for all to use, most trading is done under financial contracts that define monetary payments based on spot prices rather than under contracts that compel physical performance – not because physical contracting is more difficult than it otherwise would be, but because financial contracting is so much more flexible and efficient.

Each implementation of the basic ISO/spot market/financial contract model has been different in various ways to the others, with the later systems generally improving on the earlier in at least some respects. The principal differences are in the openness and efficiency of the ISO's spot market, the management and pricing of transmission congestion, the treatment of bilateral contracts in the settlement process, and the relationships between the ISO and the grid owner. But whatever the details of an ISO/central spot market system, the ISO always has two primary responsibilities:

1. To operate a spot market that prices (at least) energy each hour (or so<sup>6</sup>) at more-or-less what it really costs or is worth in each hour and location on the grid; and

---

<sup>5</sup> An ISO is generally an independent, non-profit entity under the control of a board representing some combination of market participants and the general public and overseen by regulators. Eventually profit-making companies will compete to provide ISO services for a fee with incentives to operate the system efficiently. An ISO could even own the grid assets – the “Transco” model – as long it remained unaffiliated with market participants, although this model creates other problems. The relative merits of the ISO/Gridco and the Transco model are beyond the scope of this paper.

<sup>6</sup> The terms “hour” and “hourly” are used here to refer to the minimum period used for dispatch and pricing. In practice, most systems use a shorter period – for example, a half-hour in England and Wales – with the trend towards ever-shorter periods. In Australia, prices are determined every five minutes but are then averaged into an hourly settlement price.

2. To manage or control by non-market means all significant network interactions and real-time effects that are not priced in the spot market.

Opponents of ISO-operated spot markets strongly attack the first of these two ISO responsibilities with the rhetorical argument that commercial trading and pricing should be left to private, competing market makers and traders, not be usurped by the monopoly ISO. But such rhetoric ignores the fact that preventing the ISO from pricing network interactions and real-time events does not eliminate these interactions and events or allow them to go unmanaged, but merely forces the ISO to manage these interactions and events by less efficient and more intrusive non-market means. If the objective really is to maximize the role of competitive market forces and minimize the extent to which the monopoly ISO determines the outcome, the ISO should operate market-clearing mechanisms that reflect network interactions and real-time events as accurately as possible.

### 3. ALTERNATIVES FOR CONTRACTING AND REAL-TIME TRADING

Despite the fact that systems based on an ISO-administered spot market and financial contracts have worked well, this model has come under intense attack, particularly in the United States and, more recently, in Britain. The alternative usually put forth involves decentralized trading of “physical” contracts or, as it called in Britain, trading outside the Pool or “TOP.” The issue of financial contracting *versus* physical contracting/TOP has become so politicized and confused that some electricity markets have been (e.g., California) or may be (e.g., England and Wales) seriously distorted in order to force physical contracting/TOP. It is worthwhile analyzing just what the real issues are in this debate.

#### 3.1 “PHYSICAL” VERSUS “FINANCIAL” CONTRACTS

Trying to define in the abstract the difference between a physical contract and a financial contract for electrical energy on a grid is an interesting metaphysical exercise. But the intense debate about physical versus financial contracting, or TOP, is about money, not metaphysics, which brings it down to earth. The only differences between physical and financial contracting that are relevant to this debate are those that make a big difference in the ultimate allocation of the monetary costs and benefits of competition in electricity.

##### 3.1.1 Ownership or Title

For most commodities, the difference between a physical and a financial contract is clear, because somebody must always “possess” the physical commodity during the time it takes to move from initial producer to ultimate consumer. A physical contract defines when and where the buyer takes title to the contract-defined amount of the physical commodity, and the new owner then pays all costs of storage, transport, insurance, loss, damage and value changes from there to the time and place of resale or final consumption. In contrast, a financial contract dealing with the same commodity has nothing to do with transfer of title to anything other than an amount of money equal (usually) to a contract quantity multiplied by the difference between some market price or price index and a contract price. But questions of title cannot be driving this debate about electricity markets.

Because electrical energy is consumed the instant it is produced, nobody can “possess” the commodity at any time between purchase and resale/consumption. Because electricity does move on the grid (although not necessarily from the seller to the buyer in a specific transaction), it is possible to define a location at which title is deemed to be transferred, such as the seller’s or the buyer’s meter or some market “hub.” Such a definition might be used to allocate between buyer and seller the system costs that depend on energy flows, such as the costs of losses, transmission congestion or ancillary services. But any such cost allocation can be specified at least as easily in a purely financial contract that says nothing about title. The intensity of the debate about financial contracting *versus* physical contracting/TOP cannot be explained by a desire to write contracts that are deemed to transfer title as an awkward way to allocate system costs.

### 3.1.2 Commercial Terms

Nor can the intensity of the debate about physical *versus* financial contracting be explained by a desire to write contracts with any specific commercial terms. Any meaningful electricity contract will specify a quantity of energy (say, in megawatt-hours or MWh) in some specified time period (usually an hour), a contract price (in \$/MWh) for that quantity, and a more-or-less definite place, e.g., “the grid in England and Wales” or “substation 17”. The contract price and quantity can be fixed or can depend on anything the parties agree – the seller’s output, the buyer’s consumption, the weather, the level of water in a reservoir, whether or not the seller or the buyer exercises its “interruptibility” rights, any price or price index, whether a particular generating plant is operating or is using high- or low-sulfur coal, etc. But once contract parties agree what will determine the quantity and price in a contract, the contract can specify either physical delivery and transfer of title or the payment of money without physical delivery and title transfer. Commercial terms cannot define the difference between physical and financial contracts.

### 3.1.3 Insulation from the Spot Price, or Trading Outside the Pool

Physical contracting or TOP is often said to allow market participants to trade energy “directly” without being forced to “trade through” the spot market or pool. Under a financial contract for differences (CFD), the seller sells and the buyer buys all its physical energy at the spot price, and then the parties make a side payment between them that depends on the contract quantity and the difference between the contract price and the spot price.<sup>7</sup> There may (or may not) be some advantage in terms of settlement risk if market participants can exchange a contract quantity directly between themselves without this quantity being priced or settled in the spot market. But any such advantage can be obtained with a simple change in settlement accounting that has nothing to do with anything “physical”.

All that is required to allow market participants to trade “directly” rather than “through the pool” is a mechanism for informing the ISO of contract quantities that the ISO will deduct from physical quantities before determining spot market payments. Such a procedure creates no inefficiencies or competitive distortions as long as payments for system costs and benefits – including the effects of grid congestion – are based on physical flows rather than on spot market quantities.<sup>8</sup> Such a settlement procedure is planned for the Ontario electricity market and could

---

<sup>7</sup> If locational pricing is used, financial transmission rights (FTRs) can and should be used to hedge the parties against changes in the locational price differential. This possibility complicates the argument but is not the critical issue. The TOP debate has raged in England and Wales despite the fact that the Pool uses the same price everywhere.

<sup>8</sup> Much of the enthusiasm for physical contracting or TOP comes from system users wanting to escape some or all system costs or uplift. But the costs of losses, congestion management, ancillary services, settlements and most administrative overheads are required for reliable system operations and have nothing to do with the spot market *per se*. Such costs should not be allocated based on contract forms or settlement details.

be implemented even in the England and Wales Pool by changing some settlement software and procedures but little else.

With this type of contract settlement, the buyer pays the entire contract amount directly to the seller and there are no spot market payments with respect to the contract quantity. If the seller has many contracts that are settled this way, its net sales (positive or negative) to the spot market are its actual physical deliveries to the grid less the sum of its contract sales. If the buyer has many contracts that are settled this way, its net purchases (positive or negative) from the spot market are its actual physical takes less the sum of its contract purchases. Any market participant can buy and sell under multiple contracts at the same time. Contract quantities can depend on anything the contracting parties agree, as long as the ISO is notified of the contract quantities – perhaps even after the event but before final settlement invoices are computed – so that the appropriate accounting entries can be made.

Market participants may want to settle their bilateral contracts in this way in order to rearrange cash flows and credit risks. The ISO's settlement system will handle less cash and may have less non-payment risk in such a "net pool" than it has in a "gross pool". But just because market participants make more payments directly between themselves does not mean that they are not imposing risks on or are not getting benefits from the ISO's clearinghouse function.<sup>9</sup> Prudential requirements and contributions to the ISO's bad-debt costs should be based on a careful analysis of actual business risks and benefits, not on superficial differences in settlement mechanics.

Contracts under which the buyer pays the seller directly may be called "physical bilateral contracts" to distinguish them from purely financial contracts, as they will be in Ontario. But this is semantics. The prices and quantities in such a "physical" contract may have nothing to do with anything physical, and a contract for differences can influence the physical actions of the parties just as well (or as poorly). The debate about physical versus financial contracting or TOP is far too intense to be about cash flows and credit risks.

### **3.1.4 Contracts for Non-Energy Physical Services or "Green" Energy**

A contract for electricity might be called "physical" if it requires the seller to produce specified amounts of energy at specified plants or in specified ways, not just to be responsible for delivering to the buyer specified quantities of commodity electricity. Perhaps the best example is a contract under which the buyer agrees to pay an above-market price for energy as long as the seller produces energy in specified environmentally preferred or "green" ways.

The physical actions required under a contract that is "physical" in this sense will usually not be things that the ISO has any reason to care about or any easy way to know about in carrying out its primary responsibilities. For example, a green energy contract for 100 kWh per month may require the seller to produce 100 kWh per month from any of a specified set of green sources,

---

<sup>9</sup> For example, a contract supplier can inform the ISO that is no longer responsible for non-paying customers, or can go bankrupt, dumping its contract customers in the ISO's lap. The ability to do this creates risks for the ISO and benefits for contract sellers.

while the ISO has no reason to care or easy way to know how energy is being produced. Administration and enforcement of such contracts can require complex monitoring of actions and tracking of transactions to assure that the buyer is getting what it pays for and that the seller is not selling the same physical action many times. The ISO can easily provide some of the required information, such as the energy output of specific plants, but cannot easily track complex chains of transactions and should not be in the business of trying to enforce compliance with private contracts.

There are some thorny issues concerning the ISO's role in the administration of contracts that are truly physical in the sense that green energy contracts are, but these issues are not what is driving the debate about physical *versus* financial contracting or TOP. The real issues in this debate concern contracts for commodity electricity, and in particular the money to be made trading such contracts in the very short term, as discussed in the next section.

### **3.2 THE CRITICAL ISSUE: SHORT-TERM TRADING ARRANGEMENTS**

There is only one difference between physical and financial contracts that can plausibly explain the intensity of debate on this issue: What happens in real time when actual performance is required. And even here there are no real differences between the contracts themselves; the real differences are in the short-term trading arrangements under which the contracts are administered.

#### **3.2.1 Contract Damages and Remedies**

A physical contract presumably specifies that the seller will deliver and buyer will take, at a specified time and place, a specific amount of physical commodity. Under contract law, if one party fails to deliver or take the contract quantity, it is in default and must pay damages to the other party. But the damaged party must also take reasonable measures to mitigate or minimize the damages. If there is a well-defined spot market price at which the damaged party can easily buy or sell the commodity, the damage can be mitigated by buying or selling in the spot market and the mitigated damage can be easily determined by reference to the spot market price.<sup>10</sup>

If contract defaults can be remedied simply by making payments based on spot market prices, a party to any contract – even one that is called, or calls itself, “physical” – should base its physical operating decisions on spot prices rather than on its contracts. When spot prices make it profitable to produce or consume more or less than contracted quantities, one or both parties should deviate from the physical performance specified in the contract to take advantage of this opportunity. The costs to the damaged party resulting from such an “economic default” will be less than the gain to the defaulting party, and the net gain can be allocated between the parties as provided in the contract.

---

<sup>10</sup> For example, if a seller fails to deliver 100 MWh of energy contracted at a price of \$25/MWh but the buyer can buy (or sell) 100 MWh at a spot price of \$30/MWh, the damage due to default is simply 100 MWh multiplied by  $(\$30/\text{MWh} - \$25/\text{MWh}) = \$5/\text{MWh}$ , or \$500.

As both an economic and a legal matter, each party to a contract should optimize its commercial position given the contractual consequences, and then pay to or receive from the other party compensation reflecting the costs caused by any deviation from the contract. When an efficient spot market can be used both to mitigate and to remedy contract defaults, contracts will tend to be written in terms of required monetary payments rather than required physical performance. There can be no significant commercial difference between physical and financial contracts, or between trading inside and trading outside the Pool, unless there is no efficient spot market or Pool.

### **3.2.2 Expanding – and Limiting – Market Choices**

With an efficient spot market, most market participants trade under financial contracts because that is the commercially sensible thing to do, not because the spot market makes it more difficult to trade under physical contracts. Two parties can always agree that neither of them will use the spot market to mitigate or remedy damages due to failure to perform physically as required in the contract.<sup>11</sup> The parties to such a contract will be no worse off absolutely than they would be if there were no spot market. But they will be at a severe relative disadvantage compared to competitors who have not artificially hamstrung themselves in this way. If physical contracts do not survive once an efficient spot market exists, it is only because they cannot compete with the more efficient and flexible financial contracts that a spot market makes possible.

Conversely, if spot trading is impossible or very costly, financial contracting is very risky and uncertain because there is no common price that can be used to mitigate and remedy damages due to failure to perform physically. Contracts are then necessarily physical because there is no viable alternative; the seller must deliver and the buyer must take the specific contracted units of the physical product at the specific time and place, or pay high damages reflecting the high real costs caused by the default. The prospect of high costs or penalties for failure to perform physically as specified in the contract will drive both buyer and seller to incur costs to reduce the probability and the effects of deviations from contract quantities.

Spot trading is inherently difficult for commodities that are difficult to standardize, such as real estate or used automobiles or specialized human skills. But economic logic and history both demonstrate that one of the key contributors to and results of economic growth is increased specialization and standardization, which leads to more efficient short-term markets and increased financial contracting. If spot trading becomes efficient enough, most contracting becomes financial. Conversely, if there is a lot of financial contracting, spot markets must be highly efficient.

---

<sup>11</sup> It may be difficult to enforce a contract of this type, because neither party will know what the other is doing in the spot market and the courts may not enforce such an uneconomic agreement if either party challenges it. The contract will hold up only if both parties abide by it voluntarily even though either of them individually, and both of them jointly, could be better off by violating it. Such mutual suicide pacts are notoriously difficult to enforce.

Just as sound money tends to make physical barter obsolete, so an efficient spot market tends to make physical contracting obsolete, and for the same reasons: Because sensible people given the option find monetary transactions and financial contracting so much easier, cheaper and more reliable than physical barter and physical contracting. If there is some perverse objective to force a lot of physical barter, there is only one sure way to do it: Prevent the emergence of, or destroy, sound money. Similarly, the only way to force a lot of physical contracting in a commodity that is easy to standardize – such as a MWh of electrical energy at a specific time and place on an electricity grid – is to prevent the emergence of, or to destroy, efficient spot markets.

Advocates of the “reforms” being attempted in the England and Wales Pool claim that the Pool-based market is inefficient because there has been so little short-term contract trading. This is confusing cause and effect. There is nothing about the Pool that makes short-term contract trading difficult; indeed, the existence of a common Pool price makes such contracting much easier than it will be when (and if) the Pool is destroyed. There would probably be more short-term contract trading in England and Wales if the generation market were more competitive and if the regional electricity companies (RECs) bore more market risk. But if market participants do not engage in frenetic short-term contract trading now it is because the Pool provides an easy, low-cost way for buyers and sellers to settle the differences between longer-term CFD positions and real-time outcomes. Eliminating the Pool would by itself accomplish nothing except force costly short-term contract trading that is otherwise not needed, for the benefit of nobody except middlemen and oligopolists.

### **3.3 CENTRALIZED VS. DECENTRALIZED SPOT TRADING IN ELECTRICITY**

The real issue in the debate about physical contracting, or TOP, is not one of contract form or settlement mechanics, but is the role of the ISO in very-short-term trading and pricing. Advocates of physical contracting/TOP assert that the ISO should not be allowed to operate a real-time spot market integrated with dispatch because this will crowd out very-short-term trading by competitive market-makers and middlemen and allow the ISO to become an unresponsive, unimaginative monopoly. The best way to prevent this is to assure that the ISO’s real-time coordination process is not an open, efficient market, so that market participants will be forced into active short-term trading in non-ISO markets. In this view, any short-term inefficiencies this creates will be more than offset by the long-term benefits of competition in the market for market-making.

Competition is certainly better than monopoly if costs and everything else are equal. It is even worth incurring somewhat higher short-term costs if this will allow competition to replace monopoly and reduce costs in the long run. Thus, if preventing the ISO from using an efficient spot market to coordinate real-time operations does not increase operational inefficiencies and transaction costs “too much”, it might be worthwhile as a way to assure a larger role for competitive marketers.

In fact, however, an ISO-administered spot market/dispatch process is much more efficient than the alternatives for inherent reasons that even the most creative marketers are unlikely to overcome. The logical conclusion is that an integrated spot market/dispatch process is a valuable

natural monopoly that should be designed and regulated as such, not impeded or destroyed just to create business opportunities for less efficient alternatives.

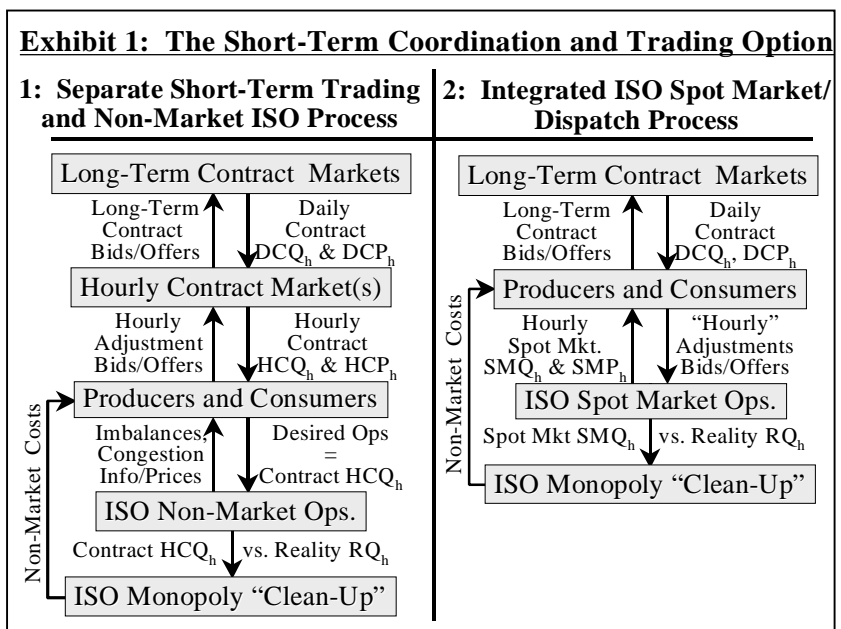
### 3.3.1 The Short-Term Coordination and Trading Options

There are two basic options for coordinating short-term system operations and trading:

**Option 1: Separated Short-Term Trading and Non-Market ISO Processes.** In this option, market participants trade contracts in diverse markets separate from the ISO, and then some time prior to operations submit to the ISO their contract-defined physical plans. The ISO uses some non-market coordination process to modify physical operations to deal with transmission constraints and with events that occur after the external markets close. The ISO imposes high penalties on deviations between contract and actual quantities (except when the deviations are requested by the ISO as part of its non-market coordination process) in order to encourage market participants to minimize such deviations through active short-term contract trading.

**Option 2: An Integrated ISO Spot Market/Dispatch Process.** As in Option 1, market participants trade contracts in diverse market processes separate from the ISO, but then some time prior to operations submit to the ISO information indicating their willingness (or unwillingness) to modify physical operations in response to prices. The ISO uses this information to determine physical operations that maximize the economic benefit to market participants as implied by their indicated willingness to modify operations, given transmission constraints and actual conditions as they develop in real time. Immediately prior to or after the fact, the ISO determines market-clearing prices that are used to settle accounts among market participants.

The basic similarities in and differences between these two approaches are illustrated in Exhibit 1. In both cases, buyers and sellers will freely negotiate and trade “long-term” contracts with terms varying from many years to a day or less. These contracts will specify prices, quantities, responsibility for losses and transmission costs, and other terms and conditions. In principle, the same commercial terms can be in the contract under either option, although in practice contracts under option 2 can be more flexible because they do not pretend to drive physical operations.



But whatever the details of the long-term contracts, each

market participant will come into each day with contract rights to take or obligations to deliver, in each hour  $h$  of the day, an amount of energy ( $DCQ_h$  in Exhibit 1) at a price ( $DCP_h$ ). The differences between the options arise only during short-term operations and trading.

### 3.3.2 Option 1: Separated Short-Term Trading and Non-Market ISO Processes

If the ISO does not operate an open, efficient spot market, each market participant knows that it will be subject to high contract costs or imbalance penalties if its final contract position does not closely match its actual physical operations. This forces each market participant actively to trade short-term contracts in non-ISO markets immediately prior to the day and during the day to try to keep its physical contract position close to its ever-changing expectations of its physical operations.

Non-ISO markets can represent physical reality only crudely, both because they cannot represent network externalities (particularly if there is more than one such market) and because they must close some time, usually hours, before reality happens. A trader in one of these markets cannot know whether its contract position will be physically feasible in real time given network constraints, the contract positions of all other traders and events after the contract market closes. But market participants will know that, whatever their final contract positions, the ISO will intervene to resolve conflicts between contracts and reality. So in determining its desired contract position, each market participant will anticipate the ISO's real-time operational process. If the ISO could use a market to price real-time externalities, such anticipation would be desirable because it would help resolve conflicts before they arise. But Option 1 does not allow the ISO to operate a market, so the ISO must use some non-market process that leaves important externalities unpriced and provides opportunities for market participants to benefit themselves individually at the expense of overall system efficiency.

Once market participants have achieved their desired hourly contract positions for the day ( $HCQ_h$  in Exhibit 1), these are submitted to the ISO as a provisional operating schedule for the day. But the ISO has important information that no individual market participant or even any one of several marketers can have, such as the contract positions of all other market participants and their cumulative impact on potential transmission constraints. In general, therefore, the ISO will discover that the contract positions or desired operational schedules, each of which may look fine from the myopic standpoint of the specific market participants involved, do not add up to a feasible solution for the system as a whole.

If the ISO is not allowed to operate a spot market, when it discovers that the market-determined solution is infeasible it can do nothing except pass that information back to market participants and tell them to try again. For example, if the ISO discovers that the market solution calls for 1,400 MW of power to flow from A to B when the transmission system can accommodate only 1,000 MW, it can announce this to the market. Market participants can then make adjustment bids/offers in the hourly (or other short-term) contract markets to try to trade into a feasible solution that involves 400 MW less power flowing from A to B without creating problems elsewhere.

When told that the system cannot handle all the flows implied by contracts as a whole, no individual trader or competitive marketer has much incentive to reduce its own use of scarce transmission in order to free it up for others. Even if some altruistic traders/marketers do try to do deals that help solve the announced problem, they cannot know what others are doing. When they all come back to the ISO with new contracts and operational schedules, the cumulative total may still be infeasible, e.g., there may still be too much power trying to get from A to B, or now there may be too much trying to go from A to C or G to H. The ISO will then have to describe the new problem and tell the market participants to go back into the now-even-shorter-term markets and try again.

This process would not necessarily converge to an efficient or even a feasible solution even if time and trading costs were no object, and will certainly often fail to produce such a solution in time given the transaction times and costs involved. Even if the interactions between the ISO and the external markets do converge to a market solution that reflects expected conditions including transmission constraints, the external market must close and the final positions must be submitted to the ISO some time before real time arrives. Thus, when hour  $h$  arrives, the ISO will invariably find conflicts between the last-submitted contracts, actual demands and actual generation availability, and the actual capability of the transmission system. The ISO must then find some way to get some market participants to produce or consume more or less than they have contracted to do.

Unless the ISO can simply issue and somehow enforce orders that require some market participants to sacrifice their commercial interests for the benefit of the system as a whole, it must use some sort of payments to induce generators to depart from their most profitable operations. The logical way to do this efficiently is for the ISO to run an auction to see which generators will modify their operations at the lowest cost. But then the ISO would be running a spot market, and the whole point of Option 1 is to prevent this. So the ISO must make a series of one-off deals with individual market participants. The cost of these monopoly-determined deals is then passed through to market participants as a whole in some sort of uplift or highly-averaged imbalance penalties that do not reflect the true scarcity value of energy or transmission capacity at particular times and places. This can “work” – but only at the cost of making the ISO a major buyer and seller for its own account in the market and distorting price signals for all market participants.

### **3.3.3 Option 2: Financial Contacts and a Centralized Spot Market**

If the ISO operates an open spot market in conjunction with the real-time dispatch, the mechanics and results of real-time operations are very different even if contracts between market participants are identical in all commercially relevant respects. Each market participant coming into the day will know that it will be able easily to “cash out” at market-clearing prices any imbalances between its final contract positions and its actual physical operations. Thus, whatever a market participant’s contract position going into the day – the  $DCQ_h$  in Exhibit 1 – or even if it has no contracts at all, it is not compelled to do any short-term contract trading either prior to the day or during the day, although it is free to do so if it chooses.

Instead of submitting its contract position or desired operating schedule to the ISO and then making adjustment offers/bids to external hourly contract markets, a market participant submits both its desired operating schedule and its hourly adjustment offers/bids to the ISO.<sup>12</sup> When the contract-determined operating schedules desired by individual market participants do not add up to a feasible solution for the system as a whole, the ISO does not just identify the current problem and tell market participants to guess again in a not-necessarily-convergent iterative process. Instead, the ISO uses the information in the voluntary bids/offers and its own knowledge of the physical system to determine the most economical way to operate the system within the physical constraints that are included in the ISO's market model.

Indeed, because the ISO has information on the prices at which market participants are willing to buy or sell energy, it is not even necessary for market participants to submit desired operating schedules. Market participants can submit offer/bid curves and the ISO can identify the opportunities for profitable trades. The market-clearing quantities and prices from the spot market –  $SMQ_h$  and  $SMP_h$  in Exhibit 1 – imply a set of spot trades among market participants that the ISO automatically identifies, schedules through its dispatch process, and settles through its settlement process.<sup>13</sup>

No real market, even one operated by the ISO as an integrated part of the dispatch process, can perfectly reflect real-time reality, so there will always be some difference between the spot-market quantities  $SMQ_h$  and actual physical operations. There will be some residual mess for the monopoly ISO to clean up outside the market, with the costs passed through to market participants somehow. But the mess and its non-market clean-up costs will be far smaller than under Option 1, because the market solution under Option 2 will more closely reflect real-time developments and grid interactions. Thus, not only does an ISO-administered spot market simplify trading for market participants, but it also reduces the ISO's role in the market.

### 3.3.4 The Natural Monopoly in Real-Time Market-Making

The fact that the ISO administers a spot market in no way prevents anybody from contracting “physically” (e.g., for “green” energy), trading short-term contracts or offering any sort of risk-management instrument. Any market participant who wants to trade very-short-term contracts to

---

<sup>12</sup> It makes no difference to operations whether a market participant submits desired operating levels with price-dependent offers to increase or decrease these levels (“inc/dec offers”) or simply submits price-dependent bids and offers; either approach defines supply or demand curves. As discussed above, market participants can submit “physical” contract quantities for “trading outside the Pool,” but these can and should be used only for settlement purposes with no effect on operations.

<sup>13</sup> It is sometimes alleged – by advocates of Pool “reform” in the UK, for example – that settling all transactions at the marginal bid or market-clearing price overpays generators or enhances generator market power relative to a “pay-as-bid” system. But a pay-as-bid system forces generators to guess at the ever-changing market-clearing price, resulting in an inefficient dispatch when they guess incorrectly. And it actually enhances generator market power, both by forcing bids to change frequently and to be significantly greater than incremental costs most of the time, and by increasing the costs and risks for smaller, undiversified players in the market.

keep its contract position closely aligned to its expected physical operations is free to do so, and if there is much demand for such trading there will be a ready supply available from competitive marketers. But no market participant is required to engage in such short-term contracting, which may add little value if a market participant can simply wait for the real-time spot market to clear any contract imbalances.

That, of course, is the critical point. It is no coincidence that the opposition to ISO-administered spot markets comes primarily from prospective market-makers and middlemen, who would naturally prefer a market in which market participants cannot function without large amounts of their services. The case is sometimes put just this bluntly.<sup>14</sup> But the more subtle claim is that the ISO has such strong advantages in operating very-short-term markets that if it is allowed to do so its monopoly position will stifle innovation in contract terms, risk-management instruments, settlement processes, etc.

There is no doubt that the ISO, as system operator, must have information and systems that give it strong natural advantages in operating very-short-term markets. The ISO must have good information about the system and every significant connected facility, must accept bids/offers to buy/sell energy for system balancing purposes, must issue operating instructions, and must assess charges and settle payments with and among market participants. Once the ISO has these capabilities, it is a relatively simple matter to add some machinery to determine market-clearing prices and to compute and manage the corresponding payments, i.e., to operate a spot market.

So independent marketers will indeed find it hard to offer very-short-term trading arrangements that can compete with an ISO-administered spot market. But that is just another way of saying that the ISO's integrated spot market/dispatch process is a valuable natural monopoly. This is a good reason to use careful, conscious, logical processes to design, govern and regulate the monopoly ISO. It is not a good reason to prohibit the ISO from operating a spot market just so that market-makers and middlemen can make money – at the expense of market participants and ultimately consumers – doing something that the ISO could do better if allowed to do so.

Once the ISO is operating its natural monopoly integrated spot market/dispatch process, there will be plenty for marketers and insurers to do. Marketers can add significant value by offering longer-term – e.g., days, weeks or months ahead – contracts that settle against the ISO's spot prices; such contracts and trading in them will be more valuable the more competitive the market is. Insurers should find a market for insurance instruments that protect market participants against the extreme price spikes that can arise in an hourly spot market. There will be many ways for enterprising, competitive marketers and others to make money operating around the ISO-administered spot market, without making it impossible for the ISO to do the job that only it can do efficiently for the ultimate benefit of final consumers.

---

<sup>14</sup> For example, in a 30 March 1999 paper submitted to OFFER, Accord Energy Ltd, a prospective electricity marketer in the UK, says: “If energy imbalance cash-out prices are perceived as likely to be commercially attractive, market participants may ... be tempted to run substantial imbalances and receive the imbalance cash-out price. ... [Unless imbalances are penalized] there is a substantial risk that bilateral trading between market participants will not flourish ...” (pp. 1 and 5)

## 4. SUCCESSES AND PROBLEMS IN ELECTRICITY MARKETS

Almost a decade after the England and Wales Pool began operating, there are over 40 system-years (over 20, not counting the 17 for Chile alone) of experience with more-or-less competitive electricity markets. This is enough to allow some of the principle successes and problems with these markets to be identified and analyzed.

This section presents a brief overview of the principal successes and problems with competitive electricity markets. Three of the most difficult problem areas – system coordination in the presence of transmission congestion, investment in generation (particularly peaking) capacity, and cost-effective competition and pricing for small consumers – are discussed, along with suggested approaches to solving problems in each area. The message is that the future of competitive electricity systems lies in developing more rather than less sophisticated spot pricing systems and using the resulting prices to support efficient financial contracting even for small consumers.

### 4.1 MANAGING TRANSMISSION CONGESTION

There is no doubt that competition in electricity markets has stimulated dramatic productivity improvements in individual business units, particularly although not exclusively in the parts of the industry that have become truly competitive. But the short-term coordination of the overall system is another matter, particularly when transmission is congested. It is very difficult to design and implement pricing systems that can replace monopoly central control of real-time operations without creating serious problems. Such problems have arisen in many cases, and various solutions have been attempted. The most effective solutions, however, are those that develop sophisticated pricing systems that internalize more of the network externalities so that the market rather than the monopoly system operator or grid owner makes more of the decisions.

#### 4.1.1 The Problem: Externalities Caused by Grid Congestion

A competitive generation market needs a price-driven dispatch so that new competitors can get into the game with more economical generation. But a price-driven dispatch does not necessarily result in the most efficient operations of a given mix of generating plants.<sup>15</sup> Indeed,

---

<sup>15</sup> The short-term dispatch process can be a complex “integer” problem that does not lend itself to price-driven solutions. For example, because starting a generating unit can be costly, it is often cheaper to start unit A that has higher marginal energy costs (in \$/MWh) but lower start-up costs (in \$/start-up) than unit B, while leaving unit B idle. Making and enforcing such decisions is relatively easy for a central dispatcher that owns all generation, but is much more difficult in a competitive market. There may be no energy price that will make it profitable both for the owner of unit A to start up and for the owner of unit B to remain idle. Either side payments must be made to get A and B to do the right thing or some inefficiency must be tolerated. Analogous situations arise in any real market and the inefficiencies are simply accepted as a fact of life. The problem is that these inefficiencies may be much larger in electricity.

the primary problem in an electricity market is often to find ways to minimize the inefficiencies that can arise when a price-driven dispatch replaces a monopoly central dispatch.

Where individual generating units are small relative to the entire market (and within sub-markets defined by transmission constraints), the short-run dispatch inefficiencies of a well-designed market are usually small relative to the over-all benefits – provided that the prices in that market reasonably reflect the realities of the transmission system. But most electricity markets use simple pricing systems that do not price transmission congestion very well or at all. The resulting externalities require the ISO to intervene in the market to assure that physical operations are consistent with transmission constraints. If the ISO's interventions are large or frequent, they become a significant determinant of commercial outcomes and much of the benefit of market pricing may be lost.

Monopoly utilities invariably say that grid congestion is not a serious problem on their system, and they are usually right in the sense that grid congestion is easy to manage for a monopolist who controls all the generation on the system. The monopoly dispatcher simply instructs some plants to produce a bit more and some to produce a bit less until the system is balanced within transmission constraints, and none of them have much reason to resist because they all work for the same boss and all costs are pooled. But things are very different on a competitive system, where affected power plants will strongly resist requests to produce more or less than the amount that maximizes their profits given the market prices.

The most common, and superficially easy, way to manage congestion is for the ISO to buy and sell energy outside the market as needed to relieve the congestion and pass the costs through to all loads in a general system charge or uplift. This can work for awhile, or where congestion really is a minor problem. But experience worldwide demonstrates that, in the absence of congestion pricing, congestion and its costs increase surprisingly quickly until something must be done about it. For example:

- Uplift costs began increasing to unexpected levels almost immediately in the England and Wales Pool, where there is no congestion pricing. The solution adopted was to give NGC (the ISO in England and Wales) financial incentives to manage congestion as a monopoly. This has “worked,” but at the cost of increasing the market role of the NGC monopoly.
- The NEPOOL (New England Power Pool) market, which has no congestion pricing, found that new generators were proposing to locate in locations that would compound congestion problems. NEPOOL tried the “simple” solution of letting the ISO prohibit new generation in inconvenient locations, but this is so discriminatory and inefficient that it has been rejected as a long-term solution.
- The PJM (Pennsylvania, New Jersey, Maryland Interconnection) market began operation with a “simple” uniform price, but when congestion arose had to cancel the market within hours because market participants responding to the resulting price incentives were making it impossible for the ISO to maintain reliability.

Elsewhere, fear of such problems is one of the principal arguments used to delay the introduction of competitive markets.

#### **4.1.2 The Logical Solution: Congestion Pricing**

The ultimate solution to the problem of grid congestion must be to create ISO-administered spot pricing systems that better reflect grid constraints, along with the associated financial transmission rights or “FTRs” that make payments based on the spot prices. An ISO-administered spot market can determine real-time prices that vary by location when transmission constraints prevent energy from flowing freely from where it is cheap to where it is valuable. When network externalities are internalized in this way, market participants can be allowed to operate, contract and trade freely in response to spot prices, with less need for the ISO to intervene in the market.

When spot energy prices vary by location to reflect transmission congestion, the price differential between locations represents the spot price of congestion between those locations. All system users, whether trading in the spot market or under bilateral contracts, should pay these congestion prices. Congestion prices can be highly volatile and unpredictable, and – because they are paid to the ISO’s settlement system and not to another market participant – cannot be hedged by a contract between market participants. But the ISO’s settlement system, which collects the rents resulting from congestion prices, can rebate these rents to the market participants holding FTRs between specific locations. This allows FTR holders to use congested transmission without paying the spot congestion prices. Because the FTRs are purely financial instruments, they can be freely traded without constraining or being constrained by physical operations, although their value as hedges may decline if actual transactions diverge too much from those implied by the FTRs.

As always, the alternative to an integrated spot market/dispatch/financial contracting process is a process based on physical contracts or rights traded in decentralized markets separated from physical operations. Most commodity markets operate in the latter way – but then most commodities do not flow on an integrated grid where network externalities are so strong and complex that a monopoly system operator is needed. Network externalities on any complex electricity grid make it virtually impossible to define physical transmission rights that will use the system fully and yet can be traded in decentralized markets.

For example, it may be possible to move 100 MW of power from location A to location B – as long as no more than 75 MW is flowing from C to D and no less than 35 MW is flowing from E to F. In this case, 100 MW of tradable A-to-B rights could be allocated to market participants; but such rights would be contingent upon all other market participants acting in such a way that the C-to-D and E-to-F flows remained in ranges that allow 100 MW to flow from A to B. If other market participants traded among themselves in a way that changed these other flows, or if somebody wanted to trade A-to-B rights for C-to-D rights, the ISO would have to decide what combinations were physically possible and which market participants should be allowed to do what. Even in concept it is hard to imagine a set of physical rights and decentralized trading arrangements that could deal with such complex externalities at all, much less rapidly and efficient enough to control real-time physical operations.

A discussion of the theory and practice of congestion pricing is beyond the scope of this paper. Suffice it to say that the theory is well developed<sup>16</sup> and that some large and complex electricity systems, most notably the PJM Interconnection, are now using congestion pricing and FTRs. Not only can this seemingly complex approach be made to “work,” but once implemented it is much easier and more logical than operating a market that ignores the complexity of grid congestion and expects the ISO to deal with it somehow.

### 4.1.3 The Trend: Locational Energy Pricing

The clear trend in electricity markets worldwide is to price congestion in the ISO-administered spot market. For example:

- Chile always has had, since 1982, a simple form of locational pricing in the spot market; other Latin American systems modeled on Chile (e.g., in Peru) have similar systems.
- Argentina has used locational pricing from the beginning of its market, in the mid-1990s.
- Norway has, since the early 1990s, established separate pricing zones within its market when transmission becomes congested.
- New Zealand has used a sophisticated nodal pricing system since the opening of its market.
- PJM (the Pennsylvania, New Jersey, Maryland Interconnection), after the virtual collapse of its uniform pricing system in the presence of congestion, introduced a sophisticated system of nodal prices and FTRs in 1998.
- The California market introduced an incomplete form of congestion pricing based on zones; problems with this system will probably force an evolution toward nodal pricing before too long.
- The states of eastern Australia are combining to form a National Electricity Market, which has different prices in each of several zones; like California, this system will probably evolve toward nodal prices eventually.

The need for locational or congestion pricing will only increase as electricity markets mature, and particularly as neighboring grids integrate as they are doing in eastern Australia. As US and European regional markets become more integrated, locational energy price that reflect transmission congestion will be inescapable. Even if the price differentials due to transmission constraints can be ignored within a small area, they certainly cannot be ignored as markets expand. It may be some years before there are fully integrated ISO-administered spot market/dispatch processes determining nodal prices in multiple states and countries, but that is surely the long-run future.

---

<sup>16</sup> The principle developer and proponent of nodal spot pricing and FTRs (or Transmission Congestion Contracts, TCCs) is Professor William Hogan of Harvard’s Kennedy School of Government.

## 4.2 INVESTMENT IN GENERATION CAPACITY

Experience with competitive electricity systems has buried the bogeyman that nobody will invest in power plants without long-term contracts. Systems based on spot markets with no long-term contracts have seen large amounts of new generation investment. Even in Latin America – Chile, Argentina, Peru, and elsewhere – where political intervention in state-owned electricity sectors had made it impossible to finance new generation, privatization, competition and a spot market resulted in a flood of new investment by international and domestic companies without long-term power purchase agreements. Similar results were seen in England and Wales, with new gas-fired capacity displacing a lot of coal-fired capacity – leading the current British government to impose a moratorium on new gas plants and to support OFFER’s proposal to eliminate the Pool.<sup>17</sup>

Although international experience clearly demonstrates that market-driven electricity systems can stimulate large amounts of generation investment without long-term contracts, it also demonstrates that if market prices do not reasonably reflect market conditions there can be either too much or too little or the wrong kind of new investment. Administratively-determined capacity payments designed to remedy capacity shortages tend to stimulate too much investment (Argentina and Peru), although the excess can turn to a deficit if conditions change (Chile). Where there is no capacity payment and no mechanism to increase spot energy prices to market-clearing levels when capacity is tight, there can be too little investment overall (Alberta) or inadequate peaking capacity (Victoria).

The most common concern related to generation investment has been how to maintain adequate peaking or reserve capacity – capacity that will be used only during a few hours or even minutes following failure of a generating unit or transmission facility, or during a few weeks of peak demand or only in dry years. This problem has arisen in Victoria, where the old monopoly had maintained a large plant that was used only during a few peak demand weeks each year and sometimes not even then. When this plant was sold to private investors, they decided that expected energy prices during these few weeks would not justify keeping the plant open and announced plans to mothball it. After a flurry of activity by government and market participants, the Victorian ISO was directed to contract with this plant to keep it open during the critical weeks and spread the contract costs across all consumers. This so-called “top-end” problem is a matter of serious concern in the Australian market generally.

---

<sup>17</sup> One of the reasons given for the British Government’s action was a belief that NGC (the Pool’s ISO) does not pay enough for the operating flexibility provided by coal plants relative to GCCs with take-or-pay gas contracts. If this is correct, it should be fixed, so that NGC pays more for flexibility outside the market. But GCCs are inherently more flexible than coal plants, so if NGC starts paying for flexibility the GCCs will renegotiate their gas contracts to take advantage of the flexibility prices, leaving coal plants about where they are or even worse off. Eliminating the Pool will not solve any problems – although it will increase the role of the NGC monopoly, perhaps making the market more susceptible to the kind of political interference seen recently.

#### 4.2.1 The Problem: “Free Rider” Network Externalities

The fundamental cause of problems with the amount and mix of generation capacity is failure to internalize an important network externality: The ability of a load to take power from the grid without a prior contract with a generator. Even if it were practical to prevent such “theft” by physical means, doing so would destroy one of the principal advantages of an integrated grid, namely the ability to move power instantaneously from where it is available to where it is needed.

The only practical way to internalize this externality without creating even worse problems is to define property rights to energy on the grid in financial rather than physical terms. The owner of energy flowing onto the grid has no physical way to prevent anybody else from taking that energy, but can and should have a property right to be paid what it is worth by whomever takes it. Such a property right can only be defined and enforced by the ISO’s settlement and pricing rules, which is one reason such rules must be mandatory and the same for all market participants, whether trading spot or under physical or financial contracts.

It is critical to realize that the problems of concern here do not go away in a market based on decentralized trading of “physical” contracts rather than an ISO-administered spot market. A “physical” contract for commodity energy (as opposed to “green” energy, for example) does not physically compel or even by itself financially motivate the seller to produce all of or the buyer to consume no more than the contract quantity. The only mechanism inducing compliance with a “physical” contract is the system of ISO-determined hourly spot prices or imbalance penalties. Even if the ISO is not allowed to operate a spot market, it must establish hourly energy prices and/or imbalance penalties and must pay for the reserves necessary to manage all within-hour events. If the ISO does not get these prices/penalties and reserve payments right, there will be either too little or too much incentive to invest in generation in general and in peaking capacity in particular. Eliminating the ISO spot market – or the Pool in England and Wales – will solve no problems and will create many.

If the ISO can determine energy prices (or imbalance penalties in a system of “physical” contracting) that approximate the market value of energy at each time and place, investors will have approximately the right market incentives to provide generation when and where it is needed. But the real value of energy on a grid can vary widely within a few seconds, e.g., after failure of a generating unit or transmission line. Energy prices accurately reflecting such effects would have to be determined every few seconds (at every location, in principle), and would have to increase to many thousands of times normal levels during critical seconds.<sup>18</sup> A market in which energy injections, withdrawals and contract imbalances are determined and priced only every hour (or even every five minutes) cannot possibly price all such effects accurately.

---

<sup>18</sup> In a market equilibrium with only an energy price, the price would have to go high enough, often enough during critical seconds to cover the full capital and operating costs of generation and/or load management facilities that operate only during those critical seconds.

Because hourly energy prices by definition cannot provide price signals concerning events within the hour, the ISO must act outside the market to pay some generators to follow system load as it varies within the hour and others to be ready to increase output quickly for short periods.<sup>19</sup> These ISO payments for ancillary services and operating reserves will be made to some kinds of generators and not to others and hence will have some effect on investment decisions. But these incentive effects of ISO payments are desirable because they help to internalize the externalities resulting from the use of hourly prices that, even if they reasonably approximate the average of the correct instantaneous prices over the hour, cannot reflect within-hour events.

In practice, hourly energy prices tend to be below the average of the “correct” instantaneous prices over the hour, particularly during critical hours when peaking capacity is needed. This is because the ISO must use a relatively simple, mechanical rule to determine a single energy price for the hour, whether for settling transactions in a spot market or for pricing/penalizing imbalances under physical contracts, and most simple pricing rules miss within-hour effects. Most pricing rules assume that nothing changes within each hour, although some of them assume that nothing changes within each (say) five-minute interval and then define the hourly price as the average of these prices over the hour. And most pricing rules base the price on easily estimated costs, such as unit running costs (as implied by generator offers), ignoring the more complex actions that impose higher costs or risks on the system, such as overloading transmission facilities, letting voltage drop (a “brownout”) or letting operating reserves fall to levels that increase the risk of system collapse.

The prices resulting from such simple pricing rules may reasonably approximate the average value of energy over an hour when there is plenty of capacity and conditions are stable, but tend to be below the average value of energy over hours when capacity is scarce and things change rapidly. By ignoring uncertainty and variations within the hour (or even within each five-minute interval) such pricing rules miss the fluctuations and risk-taking that create the highest costs for the system during critical hours. And the scarcer capacity is during an hour, the more such hourly prices will be below the true average value of energy over the hour, i.e., the larger will be the costs not internalized by the simply hourly pricing.

Hourly energy prices that are too low on average and particularly during critical hours will provide too little incentive to invest in or maintain generation capacity and particularly peaking capacity at critical times. As demand grows and capacity ages, the ISO will have to make higher payments for operating reserve outside the energy market in order to maintain system reliability. In principle the ISO can always pay enough for operating reserve to maintain reliable operations. But if hourly energy prices are too low the ISO’s bilateral reserve deals with generators (and interruptible loads) rather than hourly energy market prices will become the primary determinant of the amount and mix of generating capacity.

---

<sup>19</sup> The ISO will also have to pay for ancillary services that are not directly related to the timing issue under discussion here, such as reactive power and black-start capability.

#### 4.2.2 Logical Solution 1: Installed Capacity Requirements/Prices

There are two basic ways for the ISO to internalize, at least approximately, the externalities resulting from using a too-simple rule to determine hourly prices during critical hours. The first is to compel market participants to provide or pay for more generation capacity, and more peaking capacity in particular, than is commercially justified based on the hourly energy prices alone. For example, the ISO can impose an annual capacity requirement (CR) on each retailer, equal to that retailer's projected peak demand over the upcoming year plus a margin reflecting uncertainty about demand and generation availability. Each retailer must then demonstrate to the ISO that it has a portfolio of generating plants and capacity contracts that meets the ISO's criteria for quantity, reliability, location and operating characteristics. Retailers bidding for generation capacity and contracts to meet their CRs will provide additional income to generators, so that generating capacity and particularly peaking capacity has a source of income in addition to hourly energy prices and ancillary service payments.

Installed capacity requirements are common in power pools consisting of regulated, vertically integrated monopolies and have been adopted in the markets that have developed from such power pools in the Northeast United States. Such systems require the ISO to make many detailed judgments that directly affect the profitability of individual retailers and generating units, such as projected peak demands, the availability and maintenance of thermal generating units and of water for hydro units, transmission capacities, etc. And then the ISO must still enforce compliance with performance obligations and find some way to see that adequate capacity is actually available when and where needed. The required degree of ISO discretion and discrimination can be fundamentally inconsistent with a competitive market.

Shortening the period for which a CR is determined and enforced reduces the ISO's discretion and enforcement problems. A monthly CR is better than an annual CR, but a weekly CR is better than a monthly CR, and a daily CR is better yet. In the limit, an hourly CR could be determined for each retailer for each hour based on actual takes in that hour, met with capacity that is actually available during that hour, and enforced with a monetary penalty. Hourly trading of capacity would produce an hourly capacity price (in \$/MW/hour) that would have essentially the same effect as an hourly energy price (in \$/MWh). If the hourly enforcement penalty (in \$/MW) were zero for hours when the system had plenty of capacity and increased to some high amount as the system capacity margin declined to zero, such a capacity requirement system would be essentially equivalent to the capacity-related hourly energy price "adder" suggested below.

Instead of setting quantitative CRs for each retailer and letting the capacity market determine capacity prices, the ISO could pay fixed prices for capacity and let the market decide how much capacity to provide at that price, recovering its costs from retailers based on their loads. These two approaches would be equivalent if there were no uncertainty about market responses. In practice, however, it is usually much easier to estimate how much capacity is "needed" – which depends primarily on projected demand – than it is to estimate a price that will result in such a quantity being supplied by the market. An administratively-set capacity price will usually err on the high side and hence will tend to result in too much capacity in the long run. This has been the result of administratively-set capacity prices in Latin America and was the result in the

United States during the 1980s when regulators set the prices of “qualifying facility” contracts. Given the relative elasticities of supply and demand, it is better to set the quantity and let the market determine the price rather than the other way around.

Whether the ISO sets a fixed capacity quantity, or sets a fixed capacity price, or uses its discretion to decide how much capacity to buy depending on the price, it will be the ISO rather than the market that decides the amount and mix of capacity. The only way to reduce this ISO influence on the market outcome is to fix the basic problem – hourly energy prices that are below the average value of energy over the hour, particularly during hours when capacity is scarce.

#### 4.2.3 Logical Solution 2: An Hourly Energy Price “Adder”

Hourly energy prices determined by simple pricing rules tend to be too low when capacity is scarce because such rules ignore the many costly actions the ISO will take during such hours to avoid shedding load and, even more importantly, to avoid a massive blackout. In principle, the costs of each such action – overloading transmission, reducing voltage, reducing operating reserves at some risk to the system – could be estimated and included in the simple pricing rules along with the incremental costs of running generators or calling interruptible loads. But these costs cannot really be estimated and hence are usually ignored – even though they may be very high.

In order to deal with events within the hour, the ISO will pay generators (and load managers) for maintaining operating reserves that can be called quickly to deal with surprises. There will be some target level of operating reserves, usually expressed as a percentage of demand or as the size of the largest generating unit or transmission component that can fail suddenly, that the ISO will want to maintain if it is cheap and easy enough to do so. When there is plenty of capacity relative to demand, the price of operating reserve is small or zero, and the cost of meeting incremental demand is simply the cost of incremental generation. A simple pricing rule that ignores the effects of operating reserve levels – which most of them do – will give approximately the right result.

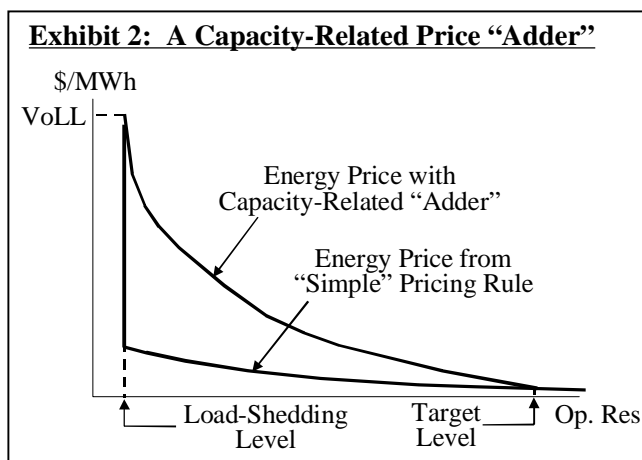
When operating reserves are at the target level and additional operating reserve is costly, the ISO may rationally decide to meet an increase in demand of, say, 1 MW, by letting operating reserves decline 1 MW below the target level. But the fact that it is rational to let this happen does not mean that it is costless to do so. If the operating reserve target is properly set and the ISO is acting rationally, the cost of letting operating reserves fall 1 MW below the target level must be approximately equal to the cost of buying an additional 1 MW of operating reserve at that point. A simple energy pricing rule that does not include this cost in determining the energy price will understate the true incremental cost of meeting demand at that time. Adding to the hourly energy price the *average* cost of operating reserve during that hour will reduce the understatement of the *incremental* energy cost, but usually not by much.

As demand increases relative to demand, the ISO may take some mix of costly actions such as reducing voltage and overloading transmission facilities, but will also let operating reserve fall even though doing so increases the risk of a very costly system collapse. When the risk of system collapse is “too high”, the ISO will start shedding load in a controlled manner. But just

before the ISO starts shedding load, it must as a logical matter be taking actions that are deemed to have costs almost as high as the cost of shedding load. If this were not the case, then the ISO should push the other, less costly actions even further before turning to the more costly option of shedding load.

Every electricity market, whether based on an ISO-administered spot market or “physical” contracts traded in decentralized markets, must have some deemed cost of shedding load that will become the spot price or imbalance penalty in extreme situations. This price, usually called the “value of lost load” or VoLL (in \$/MWh), should be very high, say 100 times the normal energy price. In concept, VoLL is the energy price at which the ISO should start shedding some load in a controlled way rather than reducing operating reserves further and increasing the risk of a system collapse. In practice, VoLL is the price or penalty that will be assessed on any consumer who takes more or any generator who delivers less than its contracted amount during critical load-shedding periods – whether the contracted amount is defined by a physical or a financial contract.

The concept of VoLL, the defects of simple pricing rules and the effects of operating reserve on real costs combine to suggest an energy price relationship similar to that illustrated in Exhibit 2. If operating reserves are at or above the target level a simple pricing rule will yield reasonable results. As operating reserves are reduced below the target level and the ISO takes increasingly costly measures that are not reflected in the simple pricing rule, the energy price should increase above that implied by the simple pricing rule.



When operating reserves are reduced to the level at which load will be shed in a controlled way to prevent a system collapse, the energy price should be VoLL. If operating reserves are reduced below the level at which load shedding begins, or are maintained at the load-shedding level but only by shedding even more load, presumably the energy price should increase above VoLL. The amount by which the “true” cost of meeting incremental load exceeds the price determined by the simple pricing rule depends on the level of operating reserves and should be added to the price resulting from the simple pricing rule.

There is no theoretically correct way to determine how the suggested capacity-related adder should vary with the level of operating reserve, beyond the general considerations outlined above. It depends on various things, such as what the simple pricing rule is and what costs it does not include. The basic point to remember is that any hourly (or even five minute) energy price is never the “right” price, but is at best only an approximation to the average of the instantaneous “right” prices over the pricing interval. A “good” pricing rule is one that produces an hourly price that approximates this average so that the energy price will provide more of the incentive for investment in generation capacity, including peaking capacity, so that the ISO’s ancillary service and operating reserve payments need provide less. If a simple pricing rule

ignores some of the costs incurred to meet demand during critical hours, it is not a very good pricing rule and should be replaced with a better one. The combination of a simple energy pricing rule and a capacity-related price adder will often be a better pricing rule than the simple rule alone.

The fact that the capacity-related price adder must be based largely on judgments that cannot be more than approximately correct is no reason to use a simple energy pricing rule that is just as judgmental but under some conditions clearly understates energy prices and hence is not even approximately correct. A relatively simple capacity-related price adder can be analyzed, debated, voted on and then incorporated in the market rules, leaving little room for ISO judgment in day-to-day pricing and requiring less direct ISO intervention in the market to correct a too-low energy price. Any pricing rule will involve judgments and approximations. Judgments that are approximately right are better than judgments that are clearly wrong.

#### **4.2.4 The Trend: No Clear Direction but Increased Recognition of the Problem**

There has been and continues to be a wide range of approaches to the problem of stimulating investment in generation capacity. Some competitive electricity systems have had so little faith in the ability of short-term markets to stimulate long-term investment that they have imposed rigid annual installed capacity requirements or administratively determined capacity payments. Others have had such blind faith in “the market” that they have not thought it necessary to worry about network externalities or how to internalize them. Only a few systems have adopted the moderate view that short-term markets can stimulate long-term investment – but only if short-term prices adequately reflect or internalize short-term system costs. These various approaches have led to various outcomes, from too much capacity to too little, and are only slowly converging to the moderate middle ground.

The England and Wales Pool has probably come the closest to the right approach, using an explicit capacity-related adder – the “LoLP×VoLL” term – in determining the day-ahead energy price. This procedure would be more logical and effective if it were combined with a more-or-less real-time market to price within-day effects.<sup>20</sup> But the fact is that the England and Wales Pool has done a better job than most in maintaining a reasonable balance between demand and capacity over an extended period of time. The current pool “reform” proposals would eliminate the LoLP×VoLL rule in favor of some yet-to-be-specified average or punitive imbalance penalty.

---

<sup>20</sup> The day-ahead Pool price for each hour is (approximately) the expected value of tomorrow’s real-time price in that hour given the day-ahead loss-of-load-probability (LoLP) and the agreed value of lost load (VoLL). The day-ahead prices are paid to all energy that is scheduled in the day-ahead market and then delivered on the day and is paid (along with the uplift) by all load on the day. The ISO (NGC) manages within-day events by making bilateral deals with individual market participants and passing the costs through to loads in uplift. The combination of the Pool prices and the ISO payments gives roughly the right price signals for generation generally, but the lack of a real-time market-clearing price tends to undercompensate generation (and load-management) capacity that can start up quickly to deal with events within the day.

The Argentine market also uses an explicit energy price adder to increase the hourly energy price (set at the beginning of each hour) as the projected operating reserve margin falls, i.e., as the probability increases that load will not be met during the hour. But this sensible energy pricing rule has been combined with a not-so-sensible “capacity payment” that is paid for energy actually generated without regard to system conditions. The result is an artificially depressed energy price (because generators must offer cheap energy in order to collect the “capacity payment”) and too much capacity (because the “capacity payment” is too high and does not decline much as excess capacity develops). The Energy Ministry has recently proposed to replace the administratively determined, energy-related “capacity payment” with a combination of annual and daily capacity markets that would determine the capacity payments necessary to maintain an administratively determined capacity margin.

Although there is still no generally accepted solution to or even definition of the problem of stimulating the right amount and mix of generation capacity, the tendency is toward development of better short-term market arrangements. Systems with annual installed capacity requirements or annual capacity payments are finding it necessary to use shorter periods to determine and/or enforce the requirements or payments. Systems with no explicit way to increase hourly energy prices when capacity is scarce are finding it necessary to pay higher short-term prices for spinning and operating reserves when capacity is scarce. Starting from either end, the logical end-point is a market that determines a combination of an hourly energy price (in \$/MWh) and an hourly capacity price (in \$/MW/hour) that clears each hour.

#### **4.3 RETAIL COMPETITION FOR SMALL CONSUMERS**

Retail competition allows final consumers to choose among alternative, competitive suppliers of electrical energy, with all suppliers using the same transmission, distribution and settlement infrastructure to assure that their customers get the physical energy they are paying for. Such competition serves three primary purposes: (1) to assure that the average wholesale generation prices are passed through to final consumers rather than captured by monopoly distribution companies or retailers; (2) to give consumers better price signals regarding the true cost of the energy they consume (or conserve) in each hour; and (3) to put competitive pressure on the non-energy services required to deliver energy to consumers, such as metering, billing and communications.

Where wholesale competition has been introduced, retail competition for the largest consumers has seldom been far behind. The largest industrial consumers usually have enough political influence, credible commercial options (e.g., self-generation or relocation) and technical capabilities (e.g., economical sophisticated metering and communication) to assure that they have access to electricity at essentially the wholesale price as soon as such a price is established. Medium-sized industrial and commercial consumers usually get access to such prices a few years after the largest consumers do.

For the smallest consumers, the situation is different. Few residential consumers are yet in the competitive market, largely because of the high cost of the sophisticated metering and complex settlement systems usually deemed necessary to allow retail competition. These consumers do not share fully in the benefits of competition at the wholesale level, both because they pay more

than wholesale prices for commodity electricity on average and because they have no opportunity to respond to spot prices. Both basic equity and efficiency could be increased significantly by reducing the costs of retail competition for small consumers to the point where such competition is cost-effective.

#### **4.3.1 The Problems: High Costs and Limited Competition for Small Consumers**

The high cost of retail competition from small consumers is largely due to informational needs. Before retailers can compete to supply individual consumers, there must be some way to hold each competitive retailer responsible for providing the amount of energy (deemed to be) taken by each of its customers in each part of the system in each hour. This requires some way to estimate each consumer's energy takes in each hour, assign each of these to the right retailer, and assure that each retailer provides the amounts of energy (including allowances for losses) taken by all of its customers in each hour in each part of the grid. This is inherently a costly process at best. But it can become unnecessarily and prohibitively costly if excessive estimation accuracy is required and if "physical" contracting is required even for small consumers.

In some cases legislators or regulators have insisted that competition be introduced for small consumers despite the large costs of doing so, apparently on faith that equally large benefits will result. In England and Wales, for example, approximately a billion US dollars were spent designing and implementing the complex metering, reconciliation and settlement system that was judged to be necessary to support retail competition for all small consumers. Whether competition for small consumers will have enough benefits to justify such high costs is far from clear. Equally importantly, there may be alternative approaches to retail competition that could produce most of the same benefits at much lower costs.

#### **4.3.2 The Logical Solution, Part 1: Correctly Define the Objectives**

The key to cost-effective retail competition is getting wholesale prices passed through to consumers at less cost. This is not just a matter of lowering the costs of sophisticated metering and information technology as far as possible and then forcing all consumers to pay these costs whatever they may be. It is more a matter of determining, or (better) establishing processes that will determine, when and where it is cost-effective to use what kinds of technology to deliver what kinds of prices. For small consumers in particular, it may be less important to deliver accurate hourly spot prices than to deliver accurate average prices without forcing the adoption of sophisticated metering and other systems that cost more than they are worth.

Retailers do not compete to supply the consumer's *actual* takes in each hour – which nobody will ever know precisely in any case – but rather to supply the consumer's *deemed* takes as determined by some estimation and allocation process. In any such process, more accuracy is better than less (if it does not cost too much) because it improves the accuracy of the cost allocation among consumers. But whatever errors the estimation process is making now it has been making for a long time. Even if the estimate of hourly takes by a consumer is not precisely

correct, all retailers will compete to minimize the cost of providing the same amount of energy at the same time and place,<sup>21</sup> and the consumer will pay on the basis of that same estimate.

Thus, accurate metering is not strictly necessary to allow retail competition to proceed. Any process that reasonably estimates hourly takes by each consumer will do, as long as the estimates do not depend on which retailer is serving that consumer. Retail competition could begin based on the load shapes and periodic meter readings that utilities have long used to set retail rates and determine bills. If and when it is cost-effective to introduce more accurate profiles or sophisticated metering, these should be introduced. But retail competition should not be delayed just because not everybody has hourly meters. Perhaps more importantly, the introduction of retail competition should not be used as an excuse to require consumers to pay for sophisticated metering that does not induce enough real-time load management to justify the high costs.

Without hourly metering, small consumers have no incentive to respond to hourly spot prices. But there is not much evidence to suggest that the great mass of small consumers will find it cost-effective to vary their loads much in response to real-time spot prices, much less that the resulting cost savings will justify much costly metering. Metering costs are falling all the time and this will help; but actually reducing or shifting physical loads usually requires investments in physical equipment that will not become cheaper just because computer chips do. If prices become "spiky" enough such investments will become cost-effective even for small consumers. But before this happens, prices may stimulate enough supply-side responses and demand-side responses by large consumers to clear hourly spot markets without much help from small consumers.

The initial emphasis in retail competition for small consumers should be on assuring that time-averaged retail electricity prices reflect time-averaged wholesale prices. Any retailer or technology vendor can try to persuade individual consumers that its sophisticated metering and control equipment will facilitate enough real-time load management to justify its costs, and any consumer so persuaded should be able to purchase and use such equipment. If consumers already buying electricity at the average wholesale price without such equipment are unimpressed by the incremental savings they could get from real-time load management, so be it. That is powerful evidence that the fancy systems are not cost effective.

The most promising metering systems may not be cost-effective if sold door-to-door, but might be if installed system-wide. If this is the case, then metering is a natural monopoly that should not be forced into the competitive market. Any vendor can try to persuade the metering utility and its regulators that a new technology is cost-effective if installed system-wide, and if the utility or regulators are persuaded the technology can be offered to or imposed on all consumers. But none of this should have anything to do with whether or not retailers should be able to

---

<sup>21</sup> If a supplier thinks that a consumer takes more or less than it actually does and arranges to deliver the incorrect amount, the difference will flow to or from some other consumer to which the wrong amount has been delivered, with the flows managed by the ISO. A system-wide reconciliation process will assure that somebody is responsible for all the energy that actually flows.

compete to sell commodity electricity. Consumers should be able to get electricity at the average wholesale price without being forced to pay for systems that are not necessary for that purpose.

Nor should consumers be forced to pay for accounting, reconciliation and settlement systems that do little except make retail competition more complex, more costly and less competitive, even if some potential suppliers want such systems. In particular, there is no consumer-oriented reason to install settlement systems complex enough to handle “physical” contracting. As discussed above, if both buyer and seller have access to the same physical spot market relatively simple financial contracts can take the place of more complex physical contracts. As discussed below, this remains true even if the buyer is a small retail consumer.

### 4.3.3 The Logical Solution, Part 2: Spot Prices and Financial Contracts

All of the above suggests a relatively simple way to create effective retail competition even for the smallest consumers: Require the monopoly local distribution company (LDC) to provide access to the spot market for all consumers who want it. Any consumer should be able to ask its local LDC to bill it or its designated retailer at the spot price (plus distribution losses) paid by the LDC for the physical electricity taken by that consumer, separate from the LDC’s charges for distribution, stranded cost recovery and other non-energy costs. The consumer’s hourly energy takes can be estimated by whatever metering and profiling/allocation procedures the LDC and its regulators agree for such a consumer.

This billing arrangement does not require (or allow) the LDC to play any commercial role other than that of a bill collector and credit-risk manager, because the LDC simply buys energy at the spot price and passes the energy and its spot price along to every consumer connected to its system. Each consumer can then decide for itself whether to buy its energy at the wholesale spot price or to enter into a contract with a retailer on some other terms – presumably terms that reduce risks, because it will be hard to beat the wholesale spot price on average.

A consumer who chooses to pay its own LDC spot-price bill can easily enter into a financial contract with somebody – even somebody who is not in the electricity business *per se* or who does not even see the consumer’s energy bill. For example, a consumer who consumes approximately 1,000 kWh per month and is billed by the LDC based on (say) Profile G can buy from its bank 1,000 kWh/month of “electricity price insurance” at 3.5 cents/kWh, or \$35/month. At the end of each month the bank will use published hourly spot prices to determine how much it would have cost to buy 1,000 kWh on Profile G, subtract \$35 from this amount, and credit (debit, if negative) the consumer’s checking account the difference. The consumer effectively buys 1,000 kWh/month for 3.5 cents/kWh, and buys or sells at the average wholesale spot price any difference between this amount and its actual monthly consumption. The bank can hedge its risks by buying CFDs in the commodity market. Such price insurance might be particularly attractive to a bank customer who already pays its LDC bill by automatic debit.

Alternatively, a consumer can contract with a retailer on whatever terms they agree, and then direct the LDC to send its spot-price bill to that retailer for payment. The retailer then pays the consumers LDC bill and bills the consumer according to the terms of their contract. In this case, the retailer is effectively buying physical energy at the consumer’s meter at the spot price and

reselling it to the consumer at the agreed contract price. The retailer can use some combination of spot purchases and financial contracts with generators (and FTRs, if locational pricing is used in the wholesale market) to obtain the delivered energy it needs to serve its customers.

Under this system, LDCs must determine, send and collect a spot-price bill for each consumer and keep track of which retailers are serving which consumers. But nobody needs to keep track of contracts between retailers and generators or to know anything about retailer/consumer contracts except the name, address and credit worthiness of the retailer serving each consumer. Small consumers who do not want to be exposed to spot prices should have no trouble finding a retailer offering a fixed price. But if there is concern that some consumers might not be served by competitive retailers and should not be forced to take the spot price, the LDC can be required to offer “default supply” based on spot prices averaged over, say, three or twelve months. And if the LDC is required to serve poor credit risks at a loss, all consumers, including those buying from a competitive retailer, should pay their fair share of these costs.

This approach to retail competition does not automatically create competition in metering and the other services often associated with retailing – but neither does it tie such services, for which competition is inherently difficult, to commodity electricity, for which competition is inherently easy. Competitive retailers and suppliers or related services should be encouraged to develop and sell services in addition to and even bundled with commodity electricity. But the best test of whether some proposed bundle really does add value to commodity electricity is to give consumers the choice between the bundle at the price asked for it and unbundled commodity electricity at the wholesale price. If competitive suppliers cannot sell their value added services to consumers who can get unbundled commodity electricity at wholesale prices without these services, then the so-called value added services are misnamed.

All advocates of competition in electricity hope that eventually most if not all of the traditionally monopoly services will become workably competitive so that monopoly supply and decision-making will no longer be necessary. But this is a hope, not yet a reality. Pure retailing should be clearly separated from the other services for which competition is difficult or impossible, and made as competitive as it can be. Then LDCs, their regulators and potential suppliers can turn to the difficult task of deciding which of these other services should be made contestable and how. Consumers should not be required to pay a toll to competitive retailers and/or suppliers of related services simply to get access to wholesale-priced energy that they should be able to get easily and cheaply directly from their LDC.

#### **4.3.4 The Trend: Slow Progress with Mixed Results**

Competition to serve small consumers is still the exception rather than the rule, even in markets where competition is thriving at the wholesale level and for large consumers. It took England and Wales nine years after establishing a competitive wholesale market to extend competition to small consumers. In Latin America, Australia and Alberta, competition for small consumers is not yet being attempted, largely because of the perceived need for sophisticated metering and complex settlement systems. Some states in the United States have allowed or mandated limited trials of “retail access”, but without creating the efficient wholesale market that would make these meaningful. New Zealand proclaimed open competition for small consumers immediately,

even before large consumers were allowed to shop, but with strict metering requirements and nonexistent regulatory enforcement that made this an empty gesture.

Norway has allowed retail competition for small consumers virtually from the beginning of its electricity market, and it is working quite well – based on energy-only meters that are sometimes read only once or a few times a year. All loads that do not have hourly meters are deemed to take a proportionate share of the net system load – total system load minus all hourly-metered loads – in each hour. Spot energy prices are relatively stable in the largely-hydro Norwegian system and energy prices are generally low, so that what is acceptable in Norway may not be elsewhere. But the experience here demonstrates how easy it all can and should be.

Ontario plans to implement a retail competition system along the lines outlined above at the same time that it begins wholesale competition in the year 2000. California did the same virtually from the beginning of its market. The legislation establishing the California market guaranteed small consumers a low price from their historic LDC – at least for a few years.<sup>22</sup> The legislated retail price has no necessary relationship to the “right” wholesale price that consumers would be charged under the scheme outlined above, but its effects has been broadly the same: Retailers have had difficulty adding any value for small consumers. One of the largest and most sophisticated retailers – Enron – declared that it could not compete with that price, withdrew from the small-consumer market and turned its attention to large consumers for whom sophisticated energy management contracts can be valuable.

Several US jurisdictions are planning to implement an overly simplified form of retail choice that reduces a consumer’s LDC bill by an arbitrary “shopping credit” if the consumer switches to a competitive retailer. If the shopping credit equals the average wholesale energy price plus any other LDC costs actually avoided when the consumer switches, this system is equivalent to the one suggested above. However, any higher shopping credit is essentially a subsidy to competitive retailers at the expense of the LDC’s shareholders and/or remaining customers (if any). Not surprisingly, competitive retailers generally support such schemes, provided that the shopping credit/subsidy is high enough.

Certainly retail competition will come for all consumers, even small ones – some day. The question is when and how, and at what cost. Some regulators and policy makers who have not yet implemented retail choice for small consumers continue to be cautious, either because they are not convinced that the benefits are worth the costs or because they are afraid of the effects on stranded costs. Others are rushing to implement schemes such as too-high shopping credits that will create a lot of retailing activity but few efficiencies, and that will exacerbate stranded cost problems. Considering how hard it is to demonstrate that competition for small consumers will provide enough real benefits to be worth its costs, caution would seem to be the wiser choice.

---

<sup>22</sup> The revenue that utilities lose due to the legislated price reduction is added to the stranded costs that are financed by state-backed bonds that consumers will pay off in future years.

## **5. CONCLUSIONS**

Competition in electricity has worked surprisingly well wherever it has been implemented with some version of an independent system operator administering an open spot market more or less integrated with physical dispatch. Competition has barely worked at all where integrated utilities have offered “open access” under rigid tariffs without an integrated spot market/dispatch system. This is not a coincidence.

An integrated spot market/dispatch process is essential for effective and efficient competition in electricity because only such a process can reasonably internalize complex network externalities. Without such an integrated spot market/dispatch process the market-determined solution is unacceptably unreliable and inefficient, forcing the necessarily monopoly system operator to intervene heavily in the market.

Where integrated spot market/dispatch processes do not yet exist, they will have to be created before competition will amount to much. Where such processes exist but are encountering problems with transmission congestion, peaking capacity, ancillary services or similar matters, it is usually because prices do not adequately reflect physical realities. The indicated solution in such cases is to develop more sophisticated pricing mechanisms that do a better job of internalizing network interactions. Such pricing usually requires and allows the definition of financial rights in scarce system capacity. The spot prices and financial rights based on such prices can and should be made available to all consumers, even the smallest ones.

Efforts to slow or even – as in England and Wales – to reverse the historic movement to an open spot market integrated with physical dispatch will create complexities and inefficiencies that benefit middlemen and oligopolists at the expense of smaller producers and final consumers. Such efforts have no logical or practical justification – assuming that the objective is to create effective and efficient competition for the ultimate benefit of consumers.