

31 March 1998

Report to the Competition Bureau of Industry Canada
Efficiency Considerations in Designing Electricity Markets

Robert Wilson¹

This report summarizes my perspective on several aspects of restructured wholesale markets for electricity. It is intended as a contribution to the Bureau's examination of design issues in jurisdictions considering restructuring.

The Introduction lays out some background and issues that motivate the subsequent discussion. The following sections consider the general architecture of wholesale markets for electricity. The first examines the choice among forms of organization, such as bilateral contracting or multilateral trading, and in the latter, the choice between a market-clearing exchange or a tight pool with centrally optimized scheduling. The second examines the transmission market in some detail, and the third examines the energy market similarly. The final two sections examine linkages among multiple markets in decentralized designs, focusing on the role of contractual commitments and the requirements for inter-market efficiency.

1. Introduction

To establish a point of departure: the current restructuring of electricity markets is consistent with the analysis by Joskow and Schmalensee in Markets for Power, 1983. They foresaw competitive markets for generation, transmission facilities operated on an open-access common-carrier basis, and retail competition among power marketers that rely on regulated utility distribution companies for delivery. Regulation of the wholesale and retail energy markets would be reduced to structural requirements and operational guidelines and monitoring, while retaining substantial regulation of the "wires" markets for transmission and distribution. These changes entail unbundling energy from T&D, thereby reversing the vertical integration of utilities.

The current issues that I address here concern mainly the organization of the wholesale markets for energy and transmission, interpreted as including ancillary services and other requirements for system reliability and security. The examination of these issues in Canada can benefit from the history of restructuring in the provinces, such as Alberta, and other countries such as Britain, Australia, New Zealand, and Norway, newly implemented designs in countries such as Spain, and current developments in several states in the U.S.² I emphasize the implications of the general principles of market design based on ideas from economics and game theory, but on some practical aspects my views are parochial because my practical experience has been mostly in California.

¹ 908 Cottrell Way, Stanford CA 94305-1012, USA. Tel: 650-493-5340. Email: rwilson@stanford.edu.

² Two useful surveys are those prepared by Putnam, Hayes and Bartlett for the Ontario Market Design Committee, and by London Economics for the California Trust for Power Industry Restructuring.

The peculiar features of the electricity industry that must be considered include temporal and stochastic variability of demands and supplies, accentuated by the non-storability of power, multiple technologies with varying sensitivities to capital and fuel costs and environmental and siting restrictions, and dependence on a reliable and secure transmission system. The economic problems include substantial non-convexities (immobility of generation and transmission facilities, scale economies in generation, non-linearities in transmission), and externalities (mainly in transmission). As regards generation these problems have eased sufficiently in recent decades to enable competitive energy markets, but they remain important considerations in designing these markets.

The criteria for selecting among market designs include efficiency over the long term, including incentives for investment in facilities for generation and transmission. However, my exposition focuses on short-term efficiency, since this is the immediate concrete problem, and it is required for long-term efficiency.

To motivate the subsequent sections, I describe three parts of the overall problem of market design. The basic design choice is the architecture of the market. There are many contending options.³ The market can be centralized or decentralized; it can be based on bilateral contracting, a centralized exchange, or a tightly controlled pool; trades can be physical or financial obligations, and they can be forward or spot contracts; the market can include financial hedges or not; the “official” market can be mandatory or optional, and encourage or discourage secondary markets. As will be evident, my opinion is that on most dimensions, the purported advantage of one extreme or the other is illusory. I favor designs that mix the two extremes to capture some of the advantages of each from parallel operations. For instance, for the three time frames of long-term, day-ahead, and real-time, there are corresponding advantages from bilateral contracting, a central exchange, and tightly controlled dispatch.

After the market architecture is established, a host of details must be specified. I do not address operational aspects here, and I refer the reader to my work for California and Alberta that elaborates the key role of procedural rules. Procedural rules must be constructed carefully to suppress gaming and promote efficiency. It is not only a matter of closing all loopholes; rather, the procedural rules must solve some basic economic problems, such as effective price discovery that enables more efficient decisions by suppliers. All this pre-supposes that the market will be sufficiently competitive to produce an efficient outcome, so if not, then further measures are required to diminish the market power of dominant incumbents and to promote entry by newcomers. The fact that I focus on the market architecture as the basic structural decision does not mean that it should be decided first; rather, parallel consideration of several designs and their implementation is useful in the early stages so that their merits can be compared in light of stakeholders’ interests.

My perspective is conditioned by my emphasis on strategic behavior. This seems paradoxical, since my aim is to construct a design that suppresses gaming or renders it ineffective in favor of greater efficiency. The principle, however, is to treat the market design as establishing a mode of competition among the traders. The key is to select a mode of competition that is most effective in realizing the potential gains from trade.

To illustrate, I describe a common fallacy. It is deceptively easy to conclude that a mandatory pool based on a centralized optimization of all generation, transmission, ancillary services, etc. –

³ A balanced and useful view of these options is presented in Issue Paper 3 of the Ontario Market Design Committee, March 1998.

as in the UK – can realize the full productive potential of the system. This view does not recognize that the schedules derived from an optimal power flow (OPF) program are no better than its inputs. In fact, suppliers can and do treat the program as a device whose outputs can be manipulated by the inputs they provide in the form of purported cost functions, availabilities, etc.⁴ Thus, the mode of competition consists of contending efforts to influence the “bottom-line” results from the OPF program, such as dispatched quantities and prices for energy, transmission, and ancillary services. In terms of economic theory: reliance on an OPF affects the form and strength of traders’ incentives at various points in the process, but it does not obviate the role of incentives. A central design problem is to identify the best locus of incentives and competitive forces.

In addition to my strategic perspective, I appreciate that traders have practical motivations that are not included in standard economic theory. For instance, suppliers are typically skeptical of designs that make their financial viability dependent on prices derived as shadow prices (Lagrange multipliers) on system constraints included in the formulation of an OPF, and centrally planned operating schedules that are several steps removed from the cost data they submit. They prefer market-clearing prices derived directly from the terms they offer, and they prefer to devise their own operating schedules to fulfill offers accepted in the market. Similarly, they are leery of intrusions by the transmission system operator (SO) into the energy markets, fearing that the SO’s extraordinary powers could bias the competitive process. I see two sources of these preferences. One is informational: submitted cost data is never sufficient to describe the full range of considerations relevant to a supplier. The other pertains to governance: the SO is usually described as the ISO, emphasizing its independence and adherence to operating standards derived from principles of power engineering, but few designs address the basic problem of incentives for the SO. For example, the SO is not liable for the financial consequences to traders of strict security standards that are motivated more by avoidance of any chance of mishap than an economic tradeoff between reliability and energy costs. Current designs rely on standards of transmission management inherited from the era when it was internalized within utilities who owned and operated transmission facilities for their native loads, but as this inheritance decays it will be useful to re-examine the issues of governance and incentives for the SO.

Radical Designs

Because the subsequent sections concentrate on designs that are close to current norms, I first mention radical designs that are excluded. One version stems from the view that the historical importance of system reliability may be less critical with the advent of computer controlled operations. For example, the airline industry has many similarities to the electricity industry but it is organized quite differently, and the reason may be that failures or errors in a transmission grid have enormous external effects throughout the system.⁵ It might be that a decade from now the best designs are more decentralized, like the airline industry, because the reliability of the

⁴ Expositions that address these issues include Mark Armstrong, Simon Cowan, and John Vickers, *Regulatory Reform: Economic Analysis and British Experience* (MIT Press, 1994, Chapter 9); Michael Einhorn (ed.), *From Regulation to Competition: New Frontiers in Electricity Markets* (Kluwer, 1994, Chapters 2-7); and Nils-Henrik von der Fehr and David Harbord, "Competition in Electricity Spot Markets: Economic Theory and International Experience" (ISBN 82-570-9166-9, Economics Dept., University of Oslo, Norway, January 1998).

⁵ The similarities include economic importance and external effects, stochastic demand, capital and fuel intensity, wastage of unused capacity (because inventories are impossible), importance for efficiency of optimal scheduling, injection (i.e., takeoff and landing) charges for use of the system, the necessity of a traffic control system for safety and reliability, the high costs of failures or errors, dependence on advanced technology, etc. This analogy is due to Severin Borenstein.

transmission system can be assured without the centralized operations inherited from vertically integrated utilities. In particular, the vulnerability of the transmission system stems presently from weak monitoring and controls on injections and withdrawals, and primitive metering devices, all of which could be eliminated by technological advances. An extreme variant imagines that the functions of the system operator could as well be managed as a franchise, provided the firm managing operations has appropriate incentives, such as liability for costs imposed on energy traders who rely on the transmission system.

Another view is that the current system designs are residues from the era of regulation in which there were inadequate incentives for product differentiation; e.g., power service differentiated by priorities or incentives for voluntary or automatic curtailment in peak periods could reduce the reliance on supply-side controls and enable more efficient investment in base-load generation facilities.

A third view is that the only unique feature of the power industry is that an optimal pricing scheme is based on congestion charges for over-demanded transmission lines, which is complicated by the implications of Kirchhoff's Laws. Organizing the entire system around this consideration seems a high cost to pay, and some argue that it would suffice to use "postage-stamp" charges for transmission, presumably differentiated by service priority, or to rely on secondary markets for trading of firm transmission rights, or even to build a transmission system sufficient to reduce congestion to a trivial minimum. This view depends on a judgment that the gains from a thoroughly optimized system for transmission and ancillary services are small compared to the gains from vigorous competition in energy markets, and in particular, avoidance of the inefficient investments (with hindsight) in generation capacity that have plagued the electricity industry over the past quarter-century.

I assume that these radical departures from current designs are not immediately relevant, if only because they imply electricity markets that are more decentralized and privately managed than is likely soon. So I focus on those design aspects that are closer to established practice.

2. Pools, Exchanges, and Bilateral Markets

The structural feature of broadest significance is the organization of the market. Among the myriad of possible forms, the ones most common in commodities markets are bilateral exchanges. Those organized as "rings" or "pits" depend on oral outcry of bids and asks (usually by brokers acting for traders), whereas others use computerized bulletin boards to post offers. Those that depend on market makers to establish prices are conducted by specialists who clear orders from a book or dealers who post bid and ask prices. Market makers are usual where it is important to sustain inter-temporal continuity of prices and reduce volatility, and typically they trade for their own accounts and maintain inventories. Market makers in the energy industries often play an important role reconciling differences among short and long term contracts, and more generally, providing a variety of contract forms and auxiliary services.

Compared to the other organizational forms discussed below, the most salient distinction of bilateral markets is the continual process of trading, with prices unique to each transaction. The experimental and empirical evidence indicates that in general bilateral markets are not less competitive or efficient than exchanges or pools. Among those with market makers, further distinctions are the "product differentiation" represented by the variety of contracts and terms tailored to individual customers, and the maintenance of some degree of price continuity.

On the other hand, bilateral markets encounter a fundamental problem maintaining efficiency in related markets for transportation or transmission. The demand for transportation is a “derived” demand; in particular, for each bilateral transaction the associated demand value for transportation to fulfill the contract is the sum of the two parties’ gains from trade in that transaction. When parties are matched somewhat randomly into pairs for bilateral transactions, their gains from trade are also random, and thus in the aggregate express inaccurately the actual demand value of transportation. When transportation is scarce or expensive, as in the case of power transmission, market makers face a substantial task in utilizing transmission facilities efficiently. They might accomplish this by aggregating transmission demands, or by brokering transmission services, but I know of no viable theory that assures the outcome is likely to be fully efficient, taking account of the inherent externalities. Thus, on matters of efficiency in transmission, faith in purely bilateral markets requires confidence in the ingenuity of market makers. This is not necessarily an argument against bilateral markets, however, since bilateral markets can operate alongside exchanges that carry more of the responsibility at the margin for insuring efficient utilization of transmission facilities.⁶ The California design includes this feature, and in Scandinavia NordPool accounts for less than 20% of the market.

Exchanges and pools offer several advantages and also bring some disadvantages compared to bilateral markets. One advantage is a central market that establishes a uniform clearing price and more accurately expresses the derived demand for transmission. The uniform clearing price has some minor potential to realize the last iota of the gains from trade, but often the motives are more practical.⁷ For a critical commodity like electricity there is also a perceived advantage in establishing an “official” exchange with minimal transactions costs, unhindered access for all traders, transparency to enable regulatory and public scrutiny, and countervailing power against the emergence of private market makers with sufficient market power to extract some portion of the potential rents. The disadvantages lie in the reliance on restrictive contract forms and inflexible procedural rules, and if the governance structure is inadequate, some potential to dictate restrictive procedures that are more convenient for administrators than traders. In addition, most pools and exchanges rely on private bilateral markets for auxiliary services such as financial contracts to hedge prices. Attempts to maintain pools and exchanges for contracts with longer terms than a day ahead have mostly failed due to lack of interest, so typically they are confined to short-forward and spot transactions.

Here I use the term exchange for a simple market clearing system. Typical examples are the exchanges in Alberta and California whose functions are confined almost entirely to establishing prices for each hour that clear the forward markets for day-ahead and hour-ahead trading. Closely related are their real-time markets conducted by the system operator, who selects among those bids offered for increments and decrements in supply and demand to manage the transmission system. Exchanges can minimize transaction costs (as evident in Alberta where transaction charges are quite small) and largely preserve traders’ prerogatives to determine their own scheduling. A disadvantage of an exchange confined solely to sales and purchases of energy is its separation from the transmission market. For example, in California the day-ahead energy market

⁶ As emphasized in the Ontario Market Design Committee’s Issue Paper 3, March 1998, this depends on thorough comparability in the treatment of bilateral contracts and exchange trades as regards charges for transmission and ancillary services.

⁷ For instance, in California the Power Exchange’s price is used to settle grandfathered contracts, and affects payments for recovery of stranded costs. Requiring the incumbent utilities to trade through the PX also makes it easier to monitor market power. In the UK initially and in Alberta still, hedging contracts used to mitigate the incentives of incumbents with substantial market power are based on the exchange price.

in the Power Exchange (PX) clears before the transmission market opens, so traders must rely on predictions about the transmission charges they will encounter later, and transmission management relies on traders' offers of incremental and decremental adjustment bids to alleviate congestion on inter-zonal lines. In some cases the exchange might be only a "pretend" market as in Alberta, where the generation and distribution subsidiaries of the major firms are so heavily hedged via contracts that the exchange price is little more than a transfer price.

I use the term pool to describe a system in which participation is mandatory and the "market" includes substantial intervention into scheduling. Pools are carried over from the operational procedures of vertically integrated utilities who entirely managed their own generation and transmission systems to serve their native loads, for which they had regulated monopolies, and in some cases, regional "tight" power pools with full control of scheduling. Typical examples today are in the U.K. and in the northeastern U.S. (New England, New York, and Pennsylvania-New Jersey-Maryland). Pools are distinguished from exchanges by the thorough integration of the energy, transmission, and ancillary services markets, and most significantly, by a centralized optimization of unit schedules that takes account of operational considerations – not just energy generation but also capacity availability, minimum generation requirements, ramping rates, etc. At the heart of such a system is a massive computer program that decides nearly all aspects of unit scheduling, usually on both a day-ahead basis and then again in real-time operations. This program is not just an OPF for energy flows but rather includes (mixed-integer nonlinear) optimization of schedules subject to system and security constraints.⁸ A price in such a system is not a market clearing price in the usual sense that it equates demand and supply; rather, it is obtained as the shadow price on a system constraint in an optimization program whose inputs include detailed operating specifications and purported cost data. Although these prices are used for settlements *ex post* as in an exchange, they do not represent prices offered by traders.

The advantage of a pool is the tight integration of all aspects of system operations, which might enable more productive efficiency, and it is invulnerable to imperfect links among the prices in a sequence of energy and transmission markets. Its disadvantages lie in the consequences of complete centralization, since it requires mandated participation and compliance with specified operating schedules. Suppliers are often reluctant to assign the prerogatives of scheduling and some are leery of prices obtained from a computer program rather than submitted bids; indeed, they may see the program as a black box whose outputs can be affected by the cost data they submit. The prices themselves are problematic since typically they include, besides energy prices, subsidy payments for capacity or availability that are more easily manipulated (as purportedly has been the case in the UK) and in any case depend on arbitrary parameters such as the assigned value of lost load and an assessed probability of lost load. Mandatory participation is a fundamental problem because it precludes development of competing markets, either exchanges or bilateral, that might prove superior or bring innovations.⁹

A point to be emphasized is that the choices among these basic organizational forms are not mutually exclusive. A system that mixes forms is feasible, such as an exchange that complements a bilateral market for forward trades, followed by real-time operations managed like a pool. One justification for a mixed system recognizes the role of timing. A pool is inherently a market for physical transactions, which is appropriate and even necessary on a short time frame such as real-

⁸ Due to the inherent complexity of this centralized optimization, such programs rely on many ad hoc techniques, so the optimization is best interpreted as an approximation.

⁹ Sources of superiority could be lower transaction costs, longer-term contracts or contracts better tailored to traders' needs, provision of auxiliary services, or differentiated products such as curtailable service or price hedges or firm transmission rights.

time operations. Exchanges and bilateral markets are essentially forward markets for financial transactions, since physical deficiencies are inconsequential and ordinarily they are settled at the subsequent spot price. Hence, the longer time frame of forward markets increases the appeal of these organizational forms.¹⁰

It is important to recognize that local preferences are important too: the New England pool is a direct extension of the familiar tight power pool that has had operating authority there for years, whereas in California the initial design based on a pool was ultimately discarded in favor of a more decentralized organization.¹¹ And of course those parties eager to profit as market makers are advocates of bilateral markets and reluctant to compete with an exchange whose transaction costs are likely to be low.

3. Transmission Management

Except in tight power pools, there is usually some separation between the markets for energy and transmission. This is partly a functional separation that isolates the complexity of transmission management from the simplicity of energy trading. It also reflects the fact that, unlike the private-good character of energy, transmission has substantial public-good aspects, pervasive externalities, and highly nonlinear behavior described by Kirchhoff's Laws. These features of transmission make the market design highly dependent on how property rights are defined.

If there were no scarcity of transmission capacity then energy markets could be conducted like other commodity markets. The fundamental problem in transmission is that real-time balancing and security requires control by a single authority that can draw on resources offered on a spot basis, or failing that, ancillary services held in reserve. Thus, real-time operations are invariably managed by a system operator (SO).¹² The design problem is therefore focused on how far to extend the authority of the SO, and in doing so, how much to rely on market processes.

One dimension is the extent of forward balancing. NordPool and California are representative of designs in which the SO clears a forward market for transmission on a day-ahead basis (and in California, also hour-ahead). Both clear on an inter-zonal basis and rely on adjustment bids (incs and decs) to alleviate congestion, imitating the procedures used by vertically integrated utilities. For the adjustment bids NordPool uses bids carried over from the energy market, whereas in California adjustment bids are voluntary and need not bear much relation to bids in the energy market.¹³ Just as there is a sequence of energy prices at which transactions in the day-ahead, hour-ahead, and real-time markets are settled, so too there is a sequence of binding usage charges for transmission that apply to these transactions. Alternative schemes defer full resolution of congestion management closer to dispatch, as in recent proposals in Alberta that would defer declarations to two hours before dispatch.

¹⁰ Issue Paper 3 of the Ontario Market Design Committee, March 1998, reflects a growing consensus that a mixed system takes best advantage of the differing features of long-term bilateral contracting, forward exchange, and real-time spot markets.

¹¹ Rebellious stakeholders in California occasionally referred to the pool design as Gosplan, alluding to the central plan in the former Soviet Union, whereas those in New England apparently view their tight pool as an obvious convenience.

¹² There is a distinction between the SO as the manager of a control area and the manager of the transmission system. When some assets or entitlements are owned by parties, such as municipal utilities, for whom the SO's transmission management is optional, the SO accepts the schedules they provide and they are immune to measures to alleviate congestion and immune to usage charges.

¹³ To avoid problems at startup, the PX initially mandates adjustment bids, but this is a temporary measure.

Even though it is the SO who conducts the day-ahead transmission market, one motive for this market is to minimize the interventions of the SO.¹⁴ That is, the aim is to enable a market for adjustment bids, seen as an extension of the day-ahead energy markets, to handle most transmission management by achieving inter-zonal balance before moving into same-day operations where the SO has tighter control on all aspects. This leaves the SO with what in California is called intra-zonal balancing, although in fact on short time frames it is managing the entire transmission system, as well as generation to follow loads. If the link between the day-ahead and real-time markets is sufficiently tight then the forward prices in the day-ahead markets can be expected to approximate the real-time prices, while providing a sufficient planning horizon for suppliers to schedule their units optimally.

The California system is also motivated substantially by the desire to enable competing forward markets for energy, so they must also compete equally in a forward market for transmission. This is carried to an extreme in the provision that the SO must retain the energy balance of each scheduling coordinator (SC) conducting an energy market; e.g., each inc/dec pair selected to alleviate congestion must come from the same SC. This runs some risk of short-run economic inefficiency because it does not assure equalization of the SCs' energy prices. This risk is viewed by some stakeholders as necessary to realize the longer-term benefits of vigorous competition among the SCs' energy markets, but it has been widely criticized because it lacks a clear economic justification. The partial remedy provided in California is allowance for inter-SC trades of adjustment bids, although due to its limited role as a pure market-clearing exchange the PX cannot easily participate in these trades.

At the other extreme from the NordPool and California forward markets are the designs that provide one form or another of transmission "rights" in the form of reservations, priorities, or insurance. These designs minimize the SO's role by auctioning reservations for most transmission capacity far in advance, such as six months or a year, and rely on trading in secondary markets to achieve an efficient reallocation for each hour. Those that provide physical rights encounter two fundamental problems. One is how to define and allocate rights in advance of the actual circumstances, such as loop flow that restricts capacity, or residual transfer capability enabled by the actual pattern of injections and withdrawals that occurs. The second is how long before dispatch to require release of a reservation if it is not scheduled, and setting penalties for noncompliance: if release is too close to dispatch then hoarding by a holder of an unused reservation could impair efficiency or enable one with market power to corner the market. For instance, if release can be deferred until after the day-ahead market then forward trades in that market can be impaired by hoarding of transmission capacity. If releases are frequent and substantial then the SO winds up managing transmission on a real-time basis, which can be precarious. And there is the practical difficulty that physical rights require the SO to monitor the allocation of rights to verify that submitted schedules conform to the entitlements owned. These considerations indicate that financial rights are preferable unless stringent controls on physical rights can ensure non-discriminatory open access to transmission.

Those systems that provide insurance or hedges issue transmission congestion contracts (TCCs) that reimburse the holder for the SO's transmission usage charge, or contracts for differences (CFDs) that achieve the same effect. In principle, private markets could provide such financial instruments, and so far the California design assumes they will, but other systems such as NY and PJM propose to rely on TCCs to allocate financially-firm transmission rights. A contentious issue

¹⁴ Another evidence of this motive is the provision that traders in the energy markets need not rely on the ancillary services acquired by the system operator, but instead can provide these themselves.

is whether holders of TCCs should be accorded priority in scheduling when there are insufficient adjustment bids to clear the forward market for transmission. Insufficiency is seen as a possible problem because traders who are fully insured by TCCs or CFDs might have reduced incentives to provide voluntary adjustment bids, so the SO might not be able to clear the day-ahead inter-zonal market with the adjustment bids it receives, implying that inter-zonal spillovers must be alleviated in real-time by attracting sufficient resources into the (supposedly intra-zonal) imbalance market. A TCC supplemented by scheduling priority is the same as a firm transmission right for most practical purposes.

All systems that rely on voluntary forward markets for adjustments to resolve congestion are vulnerable to insufficient participation by traders, with resulting spillovers into the real-time market that might be of much larger magnitudes than this market is intended to handle. Among the measures that can mitigate this problem is a high default usage charge when the adjustment market fails to clear – a price high enough to ensure that ample resources are submitted to the real-time market. An alternative is to require adjustment bids, but this can be fruitless unless there is some assurance that they reflect accurately the traders' opportunity costs; e.g., the practice in NordPool of re-using the bids in the energy market as the adjustment bids provides stronger assurance than California's design in which the submission of adjustment bids is entirely voluntary (although a high default price when the market fails provides a strong incentive to submit bids sufficient to enable the market to clear). On the other hand, the California design enables suppliers to account for their inter-temporal operating constraints via their adjustment bids. At the heart of the California design is a free-rider problem, in the sense that each trader or market-maker can take the view that it is others' responsibility to provide sufficient adjustments to clear the market for transmission. There will be preliminary evidence about whether this problem is severe when the California market begins operations in April 1998.

A major design feature of transmission markets is the price determination process, which is closely linked to the definition of property rights. As mentioned, those systems that allocate firm transmission rights or priorities (FTRs) in advance use an auction to establish initial prices that are then updated continually in secondary markets. Such systems require the auxiliary services of a SO to establish real-time prices that exhaust the residual transfer capacity of the transmission system, but the intent nevertheless is to enable secondary markets for FTRs to allocate most of the capacity. Similarly, those that provide TCCs or CFDs to hedge transmission charges still rely on a SO for real-time operations that include setting usage charges.

In its purest form, real-time congestion pricing of scarce transmission capacity sets a usage charge for each directional link in the system, or equivalently (using Kirchhoff's Laws) an injection charge at each node. The choice between these is often based on practical considerations: there may be many more links than nodes, thereby favoring nodal pricing, but perhaps only a few links are congested recurrently, in which case link pricing is simpler.¹⁵ More frequently, only a few major links or nodes are priced explicitly, and for forward markets it is sufficient to establish injection charges only for nodal hubs or for large zones, or usage charges for major inter-zonal interfaces as in NordPool and California.¹⁶ These practices have important implications for the specification of rights and hedges; e.g., secondary markets are illiquid or inactive if the FTRs or TCCs are specified in point-to-point terms rather than zone-to-zone. In principle TCCs are required for every nodal or zonal pair but in practice it suffices to consider

¹⁵ When only a few links have positive prices it is still true that nearly all nodes have nonzero injection charges.

¹⁶ In these systems the SO operator absorbs the cost of real-time intra-zonal balancing via the imbalance market.

only those nominated by traders, and then issue a subset consistent with the system capacity and security constraints. Due to loop flow, a TCC can have a negative value and require the holder to pay rather than receive a usage charge; if this is impractical then the SO must absorb the cost, whereas link prices are always nonnegative.

In a competitive market, injection or usage charges are derived from the costs of alleviating congestion, not a tariff or “postage stamp” based on embedded cost. In an optimized pool the charge represents the shadow price on capacity, but in decentralized markets it represents the difference at the margin between the cost to the SO of accepting an inc (say, of supply in an import zone) and the revenue from a dec (of supply in an export zone), or the reverse in the case of a demand inc/dec pair. For example, in a two-zone situation the usage charge for the inter-zonal interface is typically the difference in terms of \$/MWh between the most expensive inc in the import zone and the least profitable dec in the export zone, among those accepted by the SO. When the configuration is more complicated the SO uses an OPF program to select the bids that are accepted, taking account of loop flow and security constraints. Congestion pricing in this fashion is based on the principle that the transmission system is an open-access public facility in which (non-discriminatory) charges are imposed only to alleviate congestion on over-demanded interfaces. In particular, the owners of transmission assets cannot withhold capacity nor affect prices.¹⁷

Judging from systems in the U.S., where most transmission assets are privately owned, the typical flow of funds can be traced as follows. The SO sends the invoice for usage charges to the traders directly in the case of a pool, or to the management of an exchange (such as a scheduling coordinator (SC) in California) which then bills the traders, perhaps on a *pro rata* basis as in the PX. The payments to the SO are then conveyed to the holders of TCCs, if any, or to the owners of transmission assets to offset their revenue requirement for capital recovery. Revenue from auctions of FTRs or TCCs are similarly passed to the asset owners. In either case, the allocation among owners depends on an approximation of their revenue shares.

These schemes provide no incentives for owners to strengthen their transmission lines, which would reduce congestion rents, so the longer-term problem of congestion remains unsolved. Further, if the governance structure of the SO allows incumbent suppliers to veto expansion proposals, then they can foreclose opportunities to improve the competitiveness, or more accurately the contestability, of the market; indeed, it can be that all suppliers within a control area are reluctant to strengthen inter-ties that could increase imports. I know of no design presently that addresses fully the longer-term (and, due to the complex externalities and nonlinear features of transmission networks, theoretically unsolved) problem of creating incentives for efficient strengthening or expansion of the transmission system, or that collects surcharges reserved to pay for future expansion. One partial measure is that traders who build a new link to ease congestion are entitled to receive usage charges, perhaps in the form of TCCs.

Lastly, I mention a problem with transmission markets based on congestion prices. When usage charges are derived solely from the costs of alleviating congestion, traders can opt to “self-manage” congestion by curtailing their proposed power transfers sufficiently to eliminate usage charges. This is unlikely at the level of a small individual trader unless charges are imposed at the level of injection nodes or particular links. But even with large zones, market makers conducting exchanges or bilateral contracting that account for large fractions of transmission demand can

¹⁷ An exception in the U.S. is that some owners of transmission assets or grandfathered entitlements, such as municipal utilities, can opt whether to assign their capacity to the SO for transmission management. If they choose not to do so, then the SO accepts their schedules without any pricing of congestion.

self-manage in an explicit attempt to capture the congestion rents.¹⁸ The California design encourages self-management, and indeed there is no concern about who captures the rents provided congestion is alleviated one way or another. In contrast, it is fundamental to the justification for optimized pools that all congestion rents are captured via usage charges. This depends on a naïve view of incentives and strategic behavior unless market power is so dispersed that price-taking prevails. More likely, the opportunity to capture congestion rents encourages concerted efforts to capture them.¹⁹

4. The Process of Market Clearing and the Mode of Competition

The mode of competition is strongly affected by structural features of the market design. In this section I provide some examples in energy markets, and briefly, in markets for ancillary services and transmission.

Underlying these specific examples is the general view that incentive effects are not eliminated by one market design or another; rather, the form in which they are expressed depends on the specific features of the market structure. The advantage of a superior design derives from the extent to which it enables traders to express accurately the economic considerations important to them. Gaming strategies are inherent in any design that requires traders to manipulate their bids in order to take account of factors that the bid format does not allow them to express directly.

The bid format is a key factor. For example, if the market is organized to provide hourly schedules and prices, then this tends to serve the interests of demanders for whom the time of power delivery is important, and suppliers with flexibility (e.g., ponded hydro), whereas it tends to ignore the considerations of suppliers from thermal sources, who are mainly concerned with obtaining operating schedules over consecutive hours sufficient to recover the fixed costs of startup and who are unconcerned about timing *per se*. Schemes have been devised that allow demanders to bid on a time-of-day basis while suppliers bid for operating runs of various durations; prices can then be stated equivalently in terms of hourly prices for demanders and duration prices for suppliers. Similarly, for ancillary services it is usually important to distinguish between availability payments for reserving capacity and payments for delivered energy when called by the system operator. Schemes have also been devised to allow bids in terms of priorities or adjustments, such as demands that are curtailable above a specified real-time price. I bypass these more elaborate schemes here in order to focus on the basic problem of clearing an hourly market for firm energy, either forward or spot.

In energy markets there is a basic distinction between static and iterative market processes. In a static design for a pooled market each trader provides a single bid, usually in the form of a demand or supply function, with or without a separate capacity bid or a minimum revenue requirement, and perhaps in the form of a portfolio bid for multiple generation sources that is only later converted into unit schedules. The static character lies in the fact that the initial market clearing is also the final one. The theory underlying a static design is the Walrasian theory of markets, in which the market finds a price that equates stated demands and supplies. The mode of

¹⁸ This is not necessarily easy to do, since there is a significant free-rider problem engendered by each exchange's preference that others bear the greater share of the burden in curtailing their aggregate transmission demands. The game is repeated daily, however, so implicit collusion is potentially feasible.

¹⁹ Theoretical models as well as experimental results indicate that energy traders capture some portion of congestion rents, and I recall that empirical studies of the UK market confirm this prediction.

competition lies in each trader's selection of the bid function it submits – which requires substantial guesswork since others' bids are unknown when the submission is made.

If the bids are purely for hourly energy then a static design can cause problems for suppliers with fixed costs and ramping constraints because the revenue may be insufficient to cover total costs. Designs of this sort therefore provide approximate remedies: the UK provides capacity payments and Spain allows suppliers to specify a minimum revenue requirement. Without elaborating details here, my view is that these auxiliary provisions engender as many gaming problems as they solve, and in the case of capacity payments based on an assumed value of lost load, are inherently arbitrary.

An iterative market process works quite differently, and reflects the Marshallian theory of markets. As in an auction with repeated bidding, it is those traders whose bids are at the margin who contend to get their bids accepted, and in each round they can base their bids on the tentative results from previous rounds. For example, suppose that as usual a supplier's bid is submitted as a series of steps at successively higher prices. In this case a an "extra-marginal" supplier, one with a step above the market clearing price, realizes that by reducing its price for that step it can be more competitive in the next round – thereby ejecting an infra-marginal bidder who in the next round becomes extra-marginal and therefore must itself improve its offered price. Thus, Marshallian competition works by inducing competition among those bidders whose steps are actually near the margin, in contrast with Walrasian competition in which the price offered for each step must be based on a conjecture about the competitive situation in the event that step is at the margin.

Iterative processes require procedural "activity" rules to ensure serious bidding throughout (and thus reliable price discovery) and to ensure speedy convergence, but they have the advantage of avoiding *ad hoc* measures to assure bidders' fixed costs are covered.²⁰ In a day-ahead auction the key feature is that an iterative process enables "self-scheduling" in the sense that each supplier can adapt its offers in successive rounds to the observed pattern of hourly prices. With good information about the prices it can obtain in each hour, a supplier with steam plants can itself decide on which units to schedule, their start times, and their run lengths. Similarly, a supplier with ponded hydro sources can better tailor its releases to take advantage of the observed prices in peak periods. In the California PX this enables pure-energy portfolio bidding: only after the energy market clears do the portfolio bidders need to report to the system operator their unit schedules that provide the energy they sold. Instead of the detailed operating data required by the UK's static pool to run its centralized optimization program, California's decentralized design assigns authority to the suppliers to schedule their own units to meet the commitments contracted in the energy market.

These considerations are not unique to the operation of markets organized as exchanges with an hourly market clearing price that applies uniformly to all trades. Most markets for bilateral trades allow a dynamic process in which bid and ask prices are posted continually, and any posted offer can be accepted at its offered price at any time. As in an exchange using an iterative market clearing process, traders can monitor the posted prices and the prices of completed transactions to obtain good information about the prevailing pattern of prices. And because the contracts are bilateral, each party can set its own schedule to fulfill the bargain. There are also designs for

²⁰ The activity rules for the California PX are adapted from the FCC's auctions of spectrum licenses, which have been notably successful and are now used worldwide. The PX rules were tested in laboratory experiments at Caltech with good results, but they will not be implemented in the PX until late 1998, so there is presently no factual evidence on their performance in practice.

bilateral markets in which all contracts are tentative until the market clears, and then the same hourly prices apply to all completed transactions.²¹

The mode of competition for transmission is also affected by structural features of the market. At one extreme are systems that assign scheduling priority to those who hold firm transmission rights or reservations (FTRs). In these systems traders compete to acquire FTRs in the initial auction or in the secondary market, leaving the system operator with only residual responsibility for real-time balancing and security of the system. At the other extreme is the California system in which the system operator accomplishes day-ahead inter-zonal balancing by exercising options offered as adjustment bids by demanders and suppliers. Congestion on inter-zonal lines is alleviated by accepting sufficient bids for incremental generation and decremental demand in import zones, and decremental generation in export zones. Thus, in this system the transmission market is an extension of the energy market to remedy congestion by altering the location of generation.²² Intermediate designs are those in which the system operator manages transmission by setting nodal (or zonal) injection charges based on an OPF program, but traders can obtain financial insurance by acquiring TCCs or CFDs that provide hedges against the charges imposed by the system operator. In those versions in which holders of TCCs are also accorded priority in scheduling transmission, they obtain the equivalent of firm transmission rights since they are immune to the risk that transmission charges are high. In this case, traders compete for TCCs in the initial auction and in secondary markets, but only for financial insurance rather than physical rights to schedule. Of these three, the first presents some obvious problems of inefficiency and market power if FTRs can be hoarded by dominant firms, and the second might be vulnerable to insufficient adjustment bids to enable the system operator to fully alleviate congestion.²³

Ancillary services are especially sensitive to the bid format. Using spinning reserve as the example, it is clear that suppliers must be paid for capacity availability as well as energy generation. On this basis one might surmise that suppliers should bid both components, but this causes problems. The initial problem is that the system operator must evaluate such two-part bids by giving some weight (interpreted as the probability that spinning units will be called to produce) to the energy bid. But as in most multi-part bidding schemes, this is fraught with gaming problems; e.g., a bidder who thinks that a call is less probable than the weight used by the SO prefers to exaggerate the capacity bid and shrink the energy bid, and the opposite if a call is more probable. Thus the merit order of energy bids reveals less about actual costs of generation than expectations about the likelihood that spinning reserves will be activated. These incentive problems are alleviated when different procedures are used for bid evaluation and settlements. In the simplest scheme bids are accepted solely on the basis of the offered capacity price, and then settlements for energy generation are based on the system real-time energy price rather than the offered energy price.²⁴ That is, the offered energy price is interpreted only as a reserve price below which the supplier prefers not to be called. Thus, it provides a merit order for calling generation without distorting incentives. This scheme separates the competitive process into two

²¹ This design has been studied experimentally in the University of Arizona laboratory, but I have not seen a practical implementation.

²² The separation between the day-ahead energy and transmission markets in California is due to the allowance for multiple competing markets for energy, which are then reconciled in the transmission market if congestion is revealed by the schedules they submit.

²³ This is not necessarily serious on a day-ahead basis, since the main effect is spillover into real-time balancing. When the California system begins operation in the Spring of 1998 it will be clearer whether ample adjustment bids are offered.

²⁴ One qualification to this statement is that bids that would not be least cost for any real-time price are screened out before ordering the capacity bids in merit order.

parts corresponding to the two parts of the bid, one for capacity availability, and another for priority in being called to generate.

The argument is occasionally made that an energy exchange might as well augment each demand bid by the required proportion of ancillary services, or at least spinning reserve – just as is typically done for transmission losses.²⁵ This argument recognizes that on the demand side spinning reserve is a necessary complement to planned energy deliveries. It is mistaken, however, because on the supply side energy and spin are substitutes, not complements. Moreover, technologies differ considerably in their characteristics for spinning reserve; e.g., ponded hydro sources and fast-start turbines are not subject to the ramping constraints and no-load costs of steam plants, but on the other hand, thermal plants can provide spinning reserve by operating below capacity. It is better therefore to establish a separate market for spinning reserves (and curtailable loads) along with other ancillary services so that these differing characteristics can be reflected in bids.

A peculiarity of some optimized pools is payment to suppliers for capacity in addition to energy, based on so-called multi-part bids that include components for both fixed costs and incremental energy costs, with compensating charges to demanders for “uplift”. These are not payments for capacity reserved for ancillary services but rather for planned generation. This holdover from the era of regulation is unique to the electricity industry, which is the only one that does not expect suppliers to cover fixed costs, such as capital and maintenance, from the market price of its output. Although a long-run equilibrium in the industry implies prices in peak periods adequate to cover the costs of capacity idle in other periods, the motive for these payments is apparently the short-run concern that market-clearing prices for energy will be determined by incremental generation costs that will be insufficient to recover the costs of capital and O&M. Such an outcome is mainly a consequence of reliance in optimized pools on shadow prices that reflect only purported incremental costs, based on a parallel optimization of unit commitments that takes account of start-up costs, ramping constraints, and minimum generation levels, as well as the uncertainty of demand and the imputed value of lost load.²⁶ Without elaborating fully here, I am skeptical of any such payment scheme that is not tied to explicit reservation of capacity, such as for ancillary services, because I see it as an open invitation for manipulation. Designs such as those in California, Scandinavia, and Australia dispense with these payments by clearing the market for energy entirely on the basis of prices offered for delivered energy, leaving scheduling decisions to suppliers. It might indeed be that prices in California will reflect only incremental costs that are insufficient to recover the O&M costs of installed units, but if so then that signals excess capacity that in the long run should be mothballed or decommissioned.

5. Contract Commitments and Settlements

A significant dimension of market design is the character and timing of the commitments made by participants during the market process. The most important aspects of commitment are the prices on which settlements are based. Commitments are often presumed to be physical, but in

²⁵ Most systems assign to suppliers an approximate cost of losses, without attempting an exact calculation. In California, for instance, a “generation meter multiplier” is assigned to each node and updated continually to account partially for losses, and the residual is absorbed by the SO.

²⁶ It is also a consequence of relying entirely on supply-side management, taking demand as fixed and inelastic. At the very least comparable payments should be provided to demanders who accept curtailable or lower-priority service. Demand-side measures can reduce the probability and imputed value of lost load, and thereby the reliance on peaking capacity that is idle much of the year.

fact they are usually financial since a breach is remedied by charging the defaulting party the spot price of purchases or sales to make up the difference.

In a pure bid-ask market with bilateral contracts concluded continually this aspect is usually hidden by the prevailing presumption that each contract is an immediate commitment and settlement is based on the price agreed in the transaction. However, there also designs for bilateral markets in which all agreements are tentative until a final market clearing price is established that then applies uniformly to all contracts. Also, many commodities markets operate on the principle that long-term contracts are physical commitments, with settlements pegged to prices in spot markets (which often represent only a small percentage of transactions). One power market, in Finland, operates as a financial market in which prevailing prices for futures contracts provide the “signals” used by traders arranging bilateral contracts.

In markets organized as pools we can distinguish at least three forms. In an optimized tight pool in which traders submit purported costs and availabilities, a trader commits to accepting both the prices and the unit schedules obtained from the optimizing algorithm, possibly with penalties for noncompliance. Exchanges with self-scheduling can operate either as coordinating devices or as genuine price-setting mechanisms for forward contracts. Those that settle day-ahead contracts on the basis of later real-time spot prices (e.g., Alberta, Victoria) serve mainly to allocate supplies to demands on the basis of tentative clearing prices that are not binding for settlements. In an exchange there is a strong presumption in favor of using the final market clearing price even if several iterations are used to reach that conclusion. In the California PX, for instance, tentative clearing prices are established in each round, but only the final round’s prices are binding.

On standard economic grounds one might conclude that the only relevant price for allocative efficiency is the real-time spot price, and on that basis surmise that settlements should be based on this price – implying that earlier forward contracts are not binding as regards the nominal transaction price. However, this view ignores the substantial incentive effects. To motivate the subsequent discussion, I contrast the Alberta and California designs.

The design of the California PX may seem awkward at first, and indeed it is awkward in terms of the software required for settlements, since each MWh of energy might be assigned any one of several prices. In the PX’s energy market, one clearing price is financially binding for trades completed in the day-ahead forward market, another clearing price is binding in the hour-ahead forward market, and the spot price in the real-time applies to ancillary services and supplemental energy purchased by the SO. On the other hand, the advantage of this design is that traders have an incentive to bid seriously in each of the forward markets, since the trades concluded there are financially binding at the clearing price in that market.

Alberta uses the opposite design in which all settlements are made at the final spot price, calculated ex post. That this design produces incentive problems can be seen in the rules required to implement it. Traders were originally prohibited from altering their day-ahead commitments, but then pressures from suppliers led to a compromise in which each trader was allowed a single re-declaration, and lately the argument has been over whether the final time for all declarations should be moved to just two hours before dispatch. These developments reflect all suppliers’ preference to delay commitments until close to the time at which prices for settlement are established, so that uncertainty is reduced, and each supplier’s advantage from committing last so that it can take maximal advantage of the likely pattern of prices thereby revealed. The Alberta design has also invited a kind of gaming. Importers and exporters are allowed to submit multiple “virtual” declarations. They have used this opportunity to declare several alternatives on a day-

ahead basis and then to withdraw all but one shortly before dispatch in order to obtain the best terms. Of course the other traders in Alberta now want the same privilege.

My opinion is that the difficulties implementing the Alberta design are intrinsic to any design in which transactions are not financially binding at the clearing price in the market in which they are made. One can argue that a sequence of binding forward prices might sacrifice some efficiency compared to one in which settlements are based on spot prices, but my view is that this sacrifice is necessary to ensure that bids are serious in the forward markets. If viable forward markets are unnecessary, as perhaps in a purely hydro system, then spot-price settlements are sufficient, but it seems to me that justifications for forward markets also justify binding transactions at the clearing prices in these markets. One must, of course, ensure that the sequence of forward markets is sufficiently contestable to enable arbitrage that keeps forward prices in line (in expectation) with subsequent spot prices.

One should also keep in mind the range of alternatives for the form of the commitment. An important distinction is between physical and financial commitment. Bilateral markets are more dependent on physical commitments if there is not a viable spot market in which to remedy deficiencies – or at least a dealer or broker who provides the remedy. Optimized pools depend to some extent on a presumed physical commitment to the dispatch schedule, since otherwise the optimization would be a useless exercise. In other pools, however, I have yet to see a cogent argument for physical commitments, as compared to financial commitments, in forward markets. Provided those who default on prior commitments are liable for making up the difference with purchases at the spot price, the incentives for compliance are sufficient. Further, due to the considerable stochastic variation in supply and demand conditions in power markets, the flexibility allowed by purely financial commitments is superior.²⁷

The second distinction concerns the counter-party to a contract. Bilateral trades are contracts between the transacting parties, or perhaps with a dealer, whereas in an exchange or pool the counter-party to every transaction is the exchange; that is, suppliers sell to the exchange and demanders buy from it. Typically, the exchange defines standard contractual terms, and it administers the apparatus of settlements. There is no harm in this *per se*, but it encourages the growth of alternative market-makers who offer a greater variety of contractual terms and auxiliary services more closely tailored to the needs of select customers. A mandatory pool is naturally beset by pressures to remedy one or another perceived deficiency or favoritism in the rules and contracts, since invariably the pool's standard terms are inadequate to serve equally the diverse interests of a heterogeneous group of traders.

6. Multiple Markets and Inter-Market Efficiency

In its ideal form, an “optimized” pool manages everything, providing a single market for energy, transmission, ancillary services, etc. Using submitted data on availabilities, costs, and demands, and with complete data about transmission capacity, it establishes initial schedules and then supplements these based on developments in real-time. That is, it provides the services previously managed by vertically integrated utilities, or in some cases, established regional tight power pools. Here I address some of the issues that arise when this unified market is replaced by multiple markets of one form or another. I assume that transmission scheduling and real-time

²⁷ The California PX allows portfolio bids for energy, which do not require specific unit commitments. This provision provides more flexibility to suppliers with many plants, so it might be construed as favoring larger firms, but it is also true that smaller single-plant suppliers can band together to submit portfolio bids.

system control is conducted by a system operator (SO) who can draw on ancillary services and supplemental energy offers to maintain system security, balancing, and load following. I divide the discussion between parallel markets and sequential markets.

Parallel Markets

Parallel markets exist elsewhere. One is NordPool in Scandinavia, which is a “marginal” market in the sense that less than 20% of energy is traded through the exchange. This structure, consisting of a large bilateral market for long-term contracts operating in parallel with a central market for spot trades, is common in various commodity industries – prominent examples are the metals markets, where as little as 5% of trades pass through the metal exchanges even though nearly all contract prices are pegged to the spot prices.

The California design has made parallel markets a prominent issue. The debate between proponents of private bilateral markets and a pool was resolved there by allowing both. That is, in California the pool, called the Power Exchange (PX), is mandatory only for the incumbent utilities and only for a few years. Other private market makers called scheduling coordinators (SCs) can, like the PX and some large traders with direct access, submit balanced schedules for implementation by the SO (the California ISO). These private energy markets can operate in any format, as pools or bilateral contract markets or whatever they devise. The argument for the California design is that competition among alternative market designs is ultimately the best way to establish their relative merits. There are some practical reasons for establishing the PX as an official pool initially, and because the utilities are required to participate, it has a fair chance of establishing itself as the preferred market design.²⁸

Efficiency could be jeopardized by different energy prices in the various markets. If the PX remains viable, this is unlikely in the long run, since non-utility traders can trade in any market with better prices, and in any case the non-PX market makers can themselves trade in the PX to erase persistent price differentials. Admittedly this argument is asymmetric, because as a pure market-clearing mechanism the PX cannot trade in other SCs’ markets. The problem could be more substantial in the short run, since on any particular day the energy prices in the various markets might differ. The solution adopted in California is to allow inter-SC trades of adjustment bids, and in the real-time market, incs and decs that need not be paired within the same SC, and indeed for load following need not be paired at all.

The long term problem is the viability of the PX. Its role as an official market that assures open access, uniform pricing, and transparent operations would presumably not be filled by private markets. Its survival in competition with other SCs is jeopardized by its charter restriction to market clearing. For example, it cannot trade for its own account with other SCs (nor in their markets, although they can trade in the PX) to arbitrage the markets for energy and transmission. Another consideration stems from regulatory concerns. An official exchange or pool is easier to monitor and regulate. And if the market-making function for a critical commodity like electricity were dominated by private interests then new regulatory authority might be required to intervene in these markets to assure service in the public interest. This scenario has not occurred in the other basic commodity and service industries that have been deregulated, so it must rely on some aspect peculiar to the electricity industry. The presumed candidate is a market maker so

²⁸ The practical reasons include monitoring of the market power of incumbent utilities, and using the PX price to settle long-term contracts with what in the U.S. are called qualified facilities (QFs) under the 1978 PURPA regulations.

successful that it can capture monopoly rents, but my impression is that the authority of electricity industry regulators is so pervasive as to make these concerns moot at present.

Inter-Market Efficiency

A pool tries to eliminate inefficiencies by a centralized explicit optimization based on submitted cost and engineering data, some of which is monitored for accuracy. The program allocates quantities subject to system constraints, but it also obtains shadow prices used for settlements. In principle, a dual formulation could be implemented as a single market with explicit prices determined by simultaneous clearing of the markets for each of the main ingredients, such as energy, transmission, and ancillary services. Several designs have been proposed for conducting these markets simultaneously, and at least one has received some experimental testing. For example, in one version the system operator (SO) continually monitors transactions in a bilateral market based on posted bid and ask prices for energy, and then using the energy flows implied by these transactions, the SO solves a simplified dual problem that imputes shadow prices for injections at each node.

In practice, however, these markets are usually conducted in a sequence reflecting the fact that transmission demand is derived from energy transactions, and the supply is fixed. Similarly, the demand for ancillary services is nearly proportional to the demand for energy, since most system operators maintain reserves on that basis, and the supply consists mostly of residual generation capacity after accounting for the main energy transactions. Thus, the typical structure is a cascade in which the initial market is for energy, followed by a transmission market in which energy flows are adjusted to keep within the transfer capacity, then a market for ancillary services such as spinning and non-spinning reserves (for which some transfer capacity was previously set aside). These forward markets on a day-ahead (and perhaps hour-ahead) basis are followed by a real-time market in which the SO draws on supplementary offers to maintain system balancing on a short time scale, and when these are insufficient or expensive, calls on the ancillary services held in reserve.

The sequential market structure is convenient administratively and potentially as efficient as a simultaneous market. Realization of this potential depends, however, on several factors. The most obvious requirement is that the clearing prices must be tightly linked:

- The forward price for energy should be an unbiased estimator of the subsequent spot price.
- Traders transacting in the energy market should have accurate expectations about the usage charge that will be imposed later for transmission.
- Sales in the energy market should be based on accurate expectations about the opportunity cost of committing capacity there as opposed to offering it as an incremental bid in the transmission market or as reserve capacity in the ancillary services market.

The key to all three of these requirements is the accuracy, or at least the unbiasedness, of expectations about subsequent prices. Power markets are generally considered good candidates in this respect because they are repeated daily, basic energy and transmission capacity is largely fixed in the short term, and aggregate hourly demand can usually be estimated a day ahead within a few percent points – although unplanned outages and extreme weather conditions can produce larger discrepancies occasionally. In addition, that part of stochastic price variation that is insurable can be hedged via financial contracts, such as TCCs and CFDs.

Nevertheless, these favorable characteristics must be complemented with design features that provide structural support for the formation of accurate expectations. The most important is that all markets in the sequence must be easily contestable so that any significant price differences can be erased by arbitrage. Thus, systematically high prices for ancillary services should induce

higher supply bids in the energy market from suppliers who recognize that they could leave some capacity uncommitted there in order to offer it as spinning reserve. And, systematically high usage charges for transmission should attract ample incremental and decremental bids that enable the SO to reduce congestion cheaply. The most important requisite for contestability is that participation in each market is voluntary, so that traders can move from one market to another to exploit apparent price advantages.

The problem lies in the term “systematically” above, since on any particular day it could be that higher or lower prices in subsequent markets were not anticipated in earlier markets, especially the energy market. Some of these unanticipated discrepancies can be reduced by provision of informative data and predictions by the SO and by market makers; e.g., the manager of the energy market can provide reports on inter-zonal imbalances after each iteration or bilateral transaction in the energy market so that traders can better estimate the magnitude of the inter-zonal balancing that must be solved in the subsequent transmission market.

A useful structural mechanism provides corrective markets that take account of the discrepancies. The following provide some indication of how this is done in the California design.

- One example is the provision for both day-ahead markets and a repetition (typically on a smaller scale) in hour-ahead markets (actually, two hours). Thus disparities detected after the close of the day-ahead markets encourage trading in the hour-ahead markets to exploit the price differences.
- Another is that after the initial calculation of day-ahead usage charges by the SO the non-PX scheduling coordinators are allowed to trade adjustment bids before submission of their final schedules. Also, the non-PX scheduling coordinators can trade in the PX in order to arbitrage price differences between their markets.
- A third is that portfolio bids are allowed in the day-ahead energy market, so that commitments of individual generation units need not be specified until after the hourly clearing prices for energy and the interzonal power flows are established.
- A fourth is that the day-ahead energy market is conducted iteratively, which allows traders to develop some consensus about the likely pattern of energy prices across the hours of the next day, which in turn reflect expectations about transmission, ancillary services, and real-time prices.
- Lastly, the ancillary services markets are also conducted in a cascade, so that bids rejected for one service, say spinning reserve, can be carried over to compete for another service, such as non-spinning reserve.

Despite these provisions, the link between the energy and transmission markets remains the most vulnerable. An extreme occurs when the adjustment bids, if they are voluntary, are insufficient to clear the market for transmission, but more routinely it could be that usage charges are too volatile to enable reliable predictions by traders in the energy markets. Transmission pricing based solely on congestion is inherently volatile because the usage charge across an interface can be zero if capacity slightly exceeds demand, and significantly positive if the unadjusted demand slightly exceeds capacity. And other minor procedural aspects can impair predictability; e.g., if multi-zone portfolio bids are allowed then the power exchange cannot provide reliable estimates about the magnitude of the interzonal flows implied by the tentative trades during the iterative process; and prohibition against trading adjustment bids among scheduling coordinators (adopted in California as a “simplification” for the first few months to facilitate startup) can yield exaggerated usage charges because an increment from one SC cannot be matched with a

decrement from another.²⁹ For these reasons it is clear that a design priority is to strengthen the link between the day-ahead energy and transmission markets, and perhaps to adopt a design that integrates these two key markets.

7. Concluding Remarks

My examination of the architecture of wholesale electricity markets presumes that the ingredients for effective competition are present. It is important to emphasize further that market architecture is distinctly secondary in importance to market structure, in the sense of competitiveness or contestability. Monopoly power in generation, or local monopolies due to transmission constraints, can impair efficiency regardless of the market design implemented. Oligopolies are inherently more damaging to the public interest in power markets because their daily interaction offers ample opportunities for punishment strategies to police collusive arrangements, whether explicit or implicit. Thus, structural solutions to the market power of dominant incumbents are necessary.

In the same way, procedural rules are less important than architecture: no amount of fiddling with procedural rules can overcome major deficiencies in the links among the energy, transmission, and ancillary services markets. There is therefore a natural priority in the design process that starts with ensuring a competitive market structure, proceeds to the selection of the main market forums, and then concludes with the detailed issues of governance and procedures. Some procedural rules, of course, must be designed to mitigate market power and prevent collusion; e.g., it is usual to maintain the secrecy of submitted bids to thwart efforts by a collusive coalition to punish deviants.

An aspect omitted here is the role of transaction costs. This consideration affects all three stages of the design process. Procedural rules must obviously be designed to avoid unnecessary transaction costs, but it is well to realize too that a complex array of decentralized markets imposes burdens on traders, who may well prefer a simpler structure that avoids managing a complex portfolio of contracts, bids, and schedules. A simple design can also promote competition by bringing all traders together in a few markets with standardized contracts, bid formats, and trading procedures. The virtues of simplicity can be especially important in jurisdictions with few participants and small volumes of trade.

²⁹ The California design has inherent structural biases. The day-ahead transmission market relies on inc/dec pairs to balance interzonal flows, whereas the real-time market is not confined to matched pairs, and further, SCs pay the cost of interzonal balancing whereas the SO absorbs the cost of intrazonal balancing.