

Market Architecture

Robert Wilson *

Abstract: Privatization and liberalization of infrastructure industries present classic economic issues about how details of organization and procedure affect market performance. These issues are examined in the context of recent designs of auction markets, using electricity as the example. The perspective of game theory complements standard economic theory to predict effects on efficiency, incentives, and market power.

1. Introduction

A process has been underway worldwide for twenty years to privatize state enterprises and liberalize markets for the services of the infrastructure industries, including electricity, gas, telecommunications, transport and water, among others. This process is usually viewed in terms of the shift from tight regulation of vertically integrated monopolies to light regulation of functionally separated firms. This shift has been justified by changes in technology, such as diminished economies of scale reflected in the electricity industry by smaller efficient plant sizes. In airlines and trucking, contestability was viewed sufficient to limit market power, and in telecommunications, it was enforced by requiring incumbents to offer wholesale tariffs to resellers, and occasionally, access by competing carriers. Some countries have simply separated the public good embodied in the infrastructure network, such as electricity or gas transmission and rail lines, from the associated commodity or services industry. Another viewpoint emphasizes the role of unbundling to expose cross-subsidies and to improve efficiency via better pricing and incentives for greater variety of products and services. A third viewpoint, common in developing countries, sees privatization and liberalization as necessary to overcome organizational inertia with stronger incentives, and to attract private investment to serve rapidly growing demand arising from economic expansion.

The present essay examines liberalization to find lessons relevant to economic theory, and in particular, theories of market microstructure. My perspective is normative, akin to the Lange-Lerner debate of the 1930s in which the theme was how best to organize and conduct markets. The focus then was on a national economy; here it is on an industry. The normative tone reflects the increasing role of economics as an “engineering” discipline capable of providing guidance on details of market design. This role has grown as game theory and derivative theories of incentives and information have expanded economists’ tools to include methodologies for predicting how procedural aspects influence participants’ strategies and affect overall performance. Part of this toolkit

* Stanford University, Stanford CA 94305-5015 USA. Research support was provided by NSF grant SBR9511209 and the Electric Power Research Institute. I am grateful for joint work on these topics with my co-author, Hung-po Chao, to several colleagues, including Peter Cramton, Preston McAfee, John McMillan, Paul Milgrom, Shmuel Oren, Charles Plott, Alvin Roth, Jean Tirole, and Frank Wolak, and for long collaborations to Faruk Gül, Srihari Govindan, David Kreps, and John Roberts.

pertains to the standard concerns of economic policy such as productive and allocative efficiency, another part is like law in its concern for closing loopholes in procedural rules and avoiding screwups, and another concerns experimental testing *ex ante* and empirical analysis *ex post*. I intend my title to convey its double meaning in English – *architecture* as a description of the main structural features of a market, and *architecture* as the professional discipline that designs those features using a body of theory and practical skills.

The subject is too broad to address completely here, so I focus on those markets conducted as auctions in the electricity industry where I have concentrated my efforts recently. The electricity industry provides a rich context for issues of market design, and further, it illustrates the principle that a design is tightly constrained by the industry’s technology.¹ No two designs among the liberalized electricity markets are the same, so in effect an enormous experiment is underway.

Within this narrow focus, I offer mini-essays on three issues: the extent of decentralization, specification of forward and spot markets and their price determination procedures, and mitigation of market power. To set the stage, I first embed these issues within the larger context of the economic theory of markets.

1.1 Prelude

From the viewpoint of standard economic theory, wholesale markets for electricity are inherently incomplete and imperfectly competitive. Some incompleteness is inevitable because electricity is a flow (rather than a stock) that cannot be metered perfectly, and storing potential energy is expensive, but the primary source is stochastic retail demand that would be too costly to moderate via spot prices and devices for continuous control and metering. Further, flows on transmission lines are constrained continuously by operational limits and environmental factors, and ramping rates of generators are limited.

The flow aspect means that a property right cannot be assigned by title. No one owns electricity *per se*; rather, qualified market participants obtain privileges to inject or withdraw power from the transmission grid at specific locations. These privileges bring obligations to comply with technical rules and procedures for settling accounts based on metered injections and withdrawals.² Thus, all rights are reciprocal and derived from contracts. Some parties own generating plants and transmission lines but these properties are not traded in electricity markets. Various financial rights are created by contracts, such as transmission “rights,” but actually they provide only for reimbursement of the usage fees charged for transmission (with one exception in California that assures scheduling priority for a rights holder). Jurisdictions such as the U.S. require open access to the transmission system on nondiscriminatory terms, preclude owners from withholding capacity, and in some states assign control to an independent system operator.

Incompleteness would be a minor deficiency were it not that most demand values far exceed supply costs and have large stochastic and cyclical components. Given present

¹ I use the New England and California systems as prototypes of centralized and decentralized systems when describing technology. I describe mostly the supply side although in fact demand-side aspects are important; e.g., as substitutes for reserves.

² Transmission systems for rail, oil and gas meter flows but also track title to ownership of stocks. This is senseless in the case of gas, since it is a homogenous commodity.

limitations on metering and control, the compromise adopted universally is that for most retail customers the timing and quantity of power used is priced imperfectly according to crude tariffs, and in particular, no forward contract constrains the time profile of a customer's usage. That is, with few exceptions a customer has an unrestricted right to draw electricity from the grid. In fact, many jurisdictions allow suppliers a comparable unrestricted right to inject electricity, subject only to the stipulation that they are paid the spot price.

This requires strenuous efforts on the supply side to provide energy and transmission to meet expected demand, supplemented by reserves to meet contingencies. These efforts might be organized almost entirely by a continuous spot market were it not for the crucial role of transmission constraints. Because it is based on alternating current, the transmission system is highly complex and vulnerable to instability or collapse at great cost.

The chief economic consequence of the pervasive externalities and continuous requirements for balancing the transmission system is that within a short time frame (say, within an hour) a fully efficient, decentralized market solution is not feasible presently. As in other "prices versus quantities" contexts, therefore, the better alternative relies on quantity specifications. In the case of electricity, this means that real-time operations are conducted by a system operator using procedures influenced more by engineering than economic considerations, and invoking directives when markets fail.

There can be only one spot market, the one conducted by the system operator as an integral part of its technical management of the transmission system, using offers in the spot market and pre-arranged reserves to maintain stability of the system, or directives if these are insufficient. The system operator's natural monopoly of the spot market is one way the markets are imperfectly competitive; as described later, market power is also significant among participants on both the supply and demand sides.

If the spot market were complete and perfect then all forward markets could be organized around financial contracts pegged against spot prices. In fact, however, the spot market never attains this ideal. It is difficult presently to include fully such intertemporal effects as startup costs and ramping constraints, and such spatial effects as constraints on transmission lines. The end result is that the scope of the system operator's authority extends over a longer time frame to cope with the effects of the many unpriced scarce resources and coordination tasks in the system.

An important design issue is thus the scope of the system operator's authority. Implicitly, this includes its governance, regulation, and incentives, and the extent to which it relies on economic criteria. It is this issue that is addressed first.

2. The Extent of Decentralization

Restructured wholesale markets for electricity present two polar extremes, one highly centralized, the other decentralized. Examples of centralization are in Britain 1989-1999 and Pennsylvania-New Jersey-Maryland (PJM), and of decentralization, in Scandinavia and California – as well as Britain's recently proposed re-design. Centralized designs were favored initially because they imitate vertically integrated operations or inherit procedures from regional power pools, but the trend is towards decentralization as experience shows it is feasible, even if opinions differ about efficiency and system operators remain wary of reliance on markets to ensure reliability.

I discuss the strengths and weaknesses of centralized and decentralized designs as though they are dichotomous when in fact intermediate or hybrid versions are feasible and might obtain the best of both.

2.1 Centralized Systems

Two characteristics of centralized designs are a long-term relational contract among participants, and overall optimization of operational decisions. Besides specifying market rules and sanctions, the contract specifies mutual obligations, such as mandatory participation, maintenance of sufficient operable capacity, and availability for reserve duties. The key economic aspects are:

- (a) Forward (day-ahead) and real-time optimization of all generation, transmission, and reserves, including inter-temporal factors such as startup commitments and constraints on generators' ramping rates and reservoirs' potential energy.³
- (b) Pricing based on opportunity costs as measured by the shadow prices on system constraints such as the necessary equality of supply and demand in real time.

Early designs used bid formats in which suppliers specified their fixed costs of startup and these were taken into account in optimizing generation sources, but later versions require suppliers to internalize these costs.

The basic argument for centralization is that comprehensive optimization is necessary to minimize total costs of ensuring reliability and to coordinate generation, transmission, and reserves; that is, productive efficiency requires optimization. In economic terms the claimed advantage is better pricing, in the sense that shadow prices derived from constrained optimization more accurately reflect opportunity costs of scarce resources. In terms of organization, optimized operations are thought to require a "max-SO", a system operator with broad authority to manage consolidated markets. The aim is evidently a first-best solution in the sense of classic economics. In its purest form the market is conducted as a direct revelation game: participants reveal their supply costs and demand values, as well as various technical constraints, that become inputs to an algorithm.

Incentives are addressed via settlement rules that specify how financial payments are determined. Full incentive compatibility is never attempted, relying instead on competition to ensure that payment of prices derived from the shadow prices on constraints (supply = demand, transmission \leq capacity, etc.) suffices. When competition is weak, centralized systems rely mainly on strictures and sanctions to control abuses. Long-term relational contracting enables internal disciplinary powers, as in the case of New Zealand's quasi-judicial Market Surveillance Committee.

The counter-argument is that cost minimization is a fiction without stronger incentives to ensure that bids reflect actual costs. In systems like Britain where incumbents enjoy substantial market power it is sometimes obvious that bids manipulate the algorithm. From an economist's perspective, the crux is simply that optimization does not obviate participants' incentives nor mitigate market power. When competitive forces are weak, designs that ignore incentives gain little from scrupulous attention to technical constraints.

³ The typical ramping rate for a thermal generator is about 1% of rated capacity per minute, although some flexible units are designed for fast starts and higher ramping rates. Thermal generators require boilers to be heated and cooled and have minimum operating rates that are also significant constraints. Power from a hydro reservoir is nearly instantaneous, whereas nuclear units have nearly fixed operating rates.

In contrast, arguments for decentralized systems emphasize incentive effects. The theme is that the second-best solution requires maximum latitude for competitive forces to be effective, even if for practical reasons this entails some deficiencies in coordination, incomplete markets, and imperfect pricing. In terms of organization it implies a min-SO, a system operator with narrow scope to manage transmission and reserves with minimal intrusions into energy markets.

This does not preclude a max-SO in vigorously competitive situations where gains from tight coordination are relatively larger. Early preferences for a max-SO were historical residues of vertically integrated operations that survived because of naïve optimism that liberalized markets would be sufficiently competitive to suppress strategic behavior – which proved unjustified in Britain where the market power of dominant firms produced protracted struggles with the regulator, but reportedly was justified in Argentina.

Pricing and Settlements

Although efficiency is their justification, centralized systems tolerate unnecessary inefficiencies. Some are prosaic, such as reliance on a static model or a rolling horizon that ignores contingencies. Opportunities are ignored to allow suppliers to schedule their own plants using more detailed and accurate private information. Settlement procedures ignore effects on incentives and gaming. Heavy reliance is placed on penalties and sanctions when usually the optimal penalty for deviations is to charge or pay the spot price. I elaborate on two of these.

Pricing is especially vulnerable to incentive effects. An example occurs in centralized systems that settle all transactions at the real-time price. A supplier selected in the day-ahead optimization to provide a large quantity has a strong incentive to drive up the real-time price by curtailing output or exporting to contiguous regions. These adverse incentives are muted if forward transactions are settled at forward prices, with only deviations from forward contracts charged or paid the spot price. This argument for multiple settlements is vacuous in perfectly competitive markets, and some critics argue that it is generally wrong because the only economically relevant prices are spot prices since it is only in real time that supply and demand must balance physically. In fact, however, markets are imperfectly competitive, and forward markets serve the economic function of deciding irreversibly which among the operable plants will be committed to run. Recent designs of centralized systems use multiple settlements but tensions remain, as described next.

Pricing is also distorted whenever optimization is imperfect. A typical example is real-time optimization that relies on a 24-hour rolling horizon: unlike day-ahead optimization, the spot-price calculation in an hour of peak demand takes account of inter-temporal constraints on ramping down without accounting for the constraints and imputed costs of ramping up to meet the peak, so it is biased compared to the prices estimated day-ahead. The net effect is to undervalue flexible resources used to meet peak loads. A similar effect occurs whenever prices are computed periodically or averaged over longer intervals since then flexible resources are not fully compensated for short-duration price spikes.

This deficiency is one of several that might be termed model incompleteness to suggest a parallel with incomplete markets. The problem is solvable in principle by dynamic programming. The linear programming model typically used for the day-ahead

optimization could be extended to a model of linear programming under uncertainty as used in operations research; that is, extended to include contingency plans for each likely scenario of events in the hours of the next day. Such a formulation yields shadow prices that more accurately value flexible resources that can meet contingencies quickly or cheaply. In particular, it is a theoretical basis for settling forward and spot markets at their own prices even when incentive effects are irrelevant.

These two examples are indicative of a pattern in which decentralized markets are judged deficient because they are incomplete but the incompleteness of optimization models in centralized systems is not recognized.

2.2 Decentralized Systems

So how do decentralized designs fare? Operators might consider it a miracle that they work at all in systems like California and Australia with thermal generators affected by startup costs and ramping constraints (as compared say to Norway with hydro reservoirs that can vary energy generation at a moment's notice), and with nuclear units that "must

⁴ California is remarkable because of its complete reliance on voluntary participation except for designated plants that must run for local reliability. An economist's first response is more sanguine because, in effect, decentralized markets solve the dual of a primal optimization problem. The devil is in the details, however, due to two features.

- (a) Inter-temporal costs and constraints are not included explicitly and must be internalized by participants; e.g., California relies on each bidder to self-schedule its plants (startup, ramping, etc.) to generate energy sold in forward markets. This deficiency stems from simplifications necessary for implementation. Inter-temporal considerations must be internalized because the day-ahead forward market accepts separate bids for each hour of the next day and clears these 24 hourly markets independently with no allowance for cross-hour or intra-hour effects.
- (b) There is no explicit coordination of the markets for energy, transmission, and reserves. Because these markets operate in sequence and clear independently, an economist must have faith in rational expectations to believe they are nearly efficient. The matter is important because demands for transmission and reserves are essentially derived demands, and supplies are acquired by altering the initial allocation of energy production.

These features stem from limitations on the bid format and on the complexity of the market clearing process, both of which are apparently required for practical implementation. The practical solution is to provide a sequence of forward markets. This approach relies on the fact that repeated trading of a few simple contracts can approximate a complete market for contingent contracts.

2.3 A Summary Comparison

One way to obtain an overall perspective on the contrast between centralization and decentralization is to recognize that, were everything complete and perfect, they could obtain the same result. This is the primal-dual equivalence of first-best implementations

⁴ A peculiarity of electricity markets is that supplies from must-run plants are offered at non-positive bid prices, instead of subtracting the supply quantity from the demand schedule. This presentation effect led in the UK to the view that the market design favored inflexible plants compared to flexible coal-fired plants.

when vigorous competition makes the first best incentive compatible. Departures from this equivalence differ for the two designs.

Centralized designs suffer manipulations by participants with market power that raise problems because few instruments are effective counter-measures. Pricing and settlement rules sufficient for incentive compatibility are too complex to be practical, and excluded in any case by prohibitions against price differentiation of the sort used in Vickrey auctions, while punitive sanctions and penalties for abuse are inefficient to the extent they depart from prices that measure the actual marginal costs of deviations. Optimization is a fiction when detailed knowledge of participants' costs and values is replaced by submitted bids in limited formats, and impaired further when models and algorithms are insufficient to include contingency planning that recognizes the value of flexible resources.

Decentralized designs are afflicted with incomplete and weakly synchronized markets that impair coordination and contingency planning to the extent that participants' self-scheduling does not fully internalize inter-temporal considerations. Sanctions and penalties are replaced by market prices payable for deviations, which is potentially efficient, but market power can distort prices as in centralized systems. The resulting equilibrium (Nash, not Walras) has no driver to minimize total system costs.

In sum, the case for centralization is strongest when there is vigorous competition and really good optimization. The case for decentralization is strongest when tight coordination in forward markets is less important than good scheduling decisions by each participant, provided of course that a system operator manages the transmission system in real time and conducts a liquid spot market. Whether deficiencies of those optimizations used in practice are more important than incompleteness of those decentralized markets feasible in practice depends on the situation (and changes as technology changes) and depends crucially on their comparative advantages in controlling abuses of market power and stimulating competition and entry.

2.4 Hybrid Designs

One might conclude from the above that the debate over centralization is Lange-Lerner revisited.⁵ Overall optimization minimizes costs when one has good data, models, and software; and equally, decentralized markets work well when they are complete and competitive. But this primal-dual equivalence fails in practice. Markets are imperfectly competitive, severely incomplete, and poorly synchronized, so they cannot replicate an optimization exactly. On the other hand, optimization looks good because of the evident coordination, but without incentive compatibility the seeming precision of prices is an artifact. Even with accurate cost data, software limitations distort prices because inter-temporal costs and constraints are not addressed within a model that accounts for contingencies.

The resolution of these tensions lies in hybrid designs that allow more or less centralization depending on historical factors affecting organization forms. Central management of transmission and reserves by a system operator achieves most of the gains from coordination on a sort time frame, while forward markets for energy enhance

⁵ The two entries by Tadeusz Kowalik in the New Palgrave (McMillan 1987) on Lange and on Lange-Lerner, if translated from the context of a national economy to the electricity industry, indicate the continuing relevance of issues from the 1930s in micro-economic settings.

competition and promote efficiency by enabling participants to manage their own operations. The system operator intrudes increasingly as real-time approaches, ultimately conducting the spot market as the final instrument in real time for balancing supply and demand throughout the system to maintain reliability. Tolerance for allowing the SO to exercise authority over those matters not immediately germane to transmission management, especially in matters involving sanctions and penalties, depends (in former power pools) on historical familiarity and the governance structure. In this view, the market architecture is characterized mainly by the sequence of forward markets and their scope compared to the prerogatives of the system operator to integrate operations beyond merely protecting the transmission system. More complete markets and more precise pricing, especially in spot markets, eliminate unpriced externalities and reduce the SO's responsibilities.

2.5 Comparisons with other industries

The extent of centralization is resolved along similar lines in some other industries previously dominated by vertically integrated monopolies. Transmission of natural gas is managed by a system operator in two new designs (Victoria and Britain), while in the U.S. each interstate pipeline owner manages its own system, subject to regulations requiring monthly resale markets for firm transmission – and proposed regulations requiring daily auction markets for interruptible service. One can anticipate similar designs in telecommunications where each system owner may eventually offer spot markets for spectrum. In the U.S., rail companies conduct forward markets for grain cars to serve shippers in the summer harvest months. In a few countries, rail lines are centrally scheduled by a system operator based on auction markets for access from competing owners of rolling stock. Among the other transport industries, trucking and shipping are seemingly immune to network constraints, but a market for conveying empty containers to shippers may emerge. Airlines were the obvious exception, due to their anomalous retention of extraordinary powers to price discriminate and exclude resale, but primitive auction markets made inroads recently by using the market-maker as a broker to shop the traveler's bid among the airlines, and as usual, by excluding refunds and resale. This is a harbinger of more retail auction markets implemented by electronic commerce.

3. Microstructure of Forward and Spot Markets

This section examines the sequence of forward and spot markets in an electricity system, and the connections among the markets for energy, transmission, and reserves. We begin with the spot market where all aspects are consolidated, and then work backward through the forward markets.

3.1 The Spot Market

First a brief technical summary of real-time operations. As mentioned, electricity is a flow so operations are continuous. The spot market approximates continuous operation by revising prices every 5 or 10 minutes, although imperfect metering and software limitations often require settlements on a coarser time frame, such as hourly using the average price within the hour.

Because imbalances can injure or destabilize transmission links, electrical systems require continuous balancing of demand and supply. Balancing is rendered more difficult

by limited or expensive storage of potential energy in reservoirs, and for historical reasons, few storage devices (such as batteries) and backup generators at customers' sites. In all designs, a system operator (SO) balances the system continuously using offers submitted to the spot market and previously acquired options for several categories of reserves.⁶ Momentary imbalances are detected and corrected automatically by the first reserve category, called regulation, which is provided by dispersed generators equipped with speed controls that respond to frequency sensors. As regulation capacity nears exhaustion its role is replaced, with first preference given to offers in the spot market, then options on operating and replacement reserves. Operating reserves include spinning and non-spinning reserves with response times of 10 to 30 minutes. Replacement reserves (60 minutes) are activated to sustain the required margin ($\approx 7\%$) of operating reserves as spinning and then non-spinning units are called.⁷ Within each category the options are mostly used in "merit order" according to marginal cost as bid, but an option is invoked out of merit order when needed to remedy an imbalance at a particular node of the transmission system. A further category called reliability-must-run is contracted long-term and scheduled in advance to ensure voltage support at key locations.

Each category is further divided between sources of incremental and decremental energy. Thus, growing demand is met by invoking "inc" supply options, and declining demand is met by invoking "dec" supply options. It is easiest to interpret a supply inc as an offer to increase output at a price payable to the supplier, and a dec as an offer to decrease output at a price payable by the supplier; that is, a dec enables the supplier to purchase energy from the SO to replace output commitments contracted previously in the forward markets. Thus, in a stable situation the unused supply incs in merit order represent the extramarginal segment of the short-run supply curve at prices above the current spot price, and the unused decs represent the inframarginal segment at prices below the current spot price. Incs and decs from demanders have the opposite interpretations.

In economic terms, the end result is a continually adjusted real-time price for energy, plus additional transmission costs (absorbed by the SO) for options exercised out of merit order to maintain reliability. In practice, this system-wide real-time price is defined as the highest price among those options exercised in merit order. It might seem that such a market exemplifies the ideal studied by theorists, but practical aspects intrude.

In a fully centralized system, none of the options listed above is entirely voluntary and the SO has full control of real-time dispatch: typically a supplier must bid all its operable capacity in the day-ahead market and accept assignments to reserve status; indeed, every operating generator's incs and decs are included in the merit order even if not assigned reserve status. Further, the actual real-time dispatch is re-optimized every few minutes based on predicted demand over a rolling horizon as long as 24 hours to take account of ramping constraints. Thus, as mentioned in section 2, spot prices include anticipated inter-temporal effects.

Fully decentralized systems operate differently. From the SO's viewpoint, reliability seems precarious because participation in forward markets for reserves and the spot

⁶ I ignore other factors here, such as provision of reactive energy.

⁷ A reserve unit is non-spinning if it is not synchronized with the transmission grid. Short-response non-spinning reserves are typically provided by hydro units and combustion turbines. The regional reliability councils have differing reserve requirements, and hydro resources are allowed smaller reserve margins.

market is voluntary – insufficient offers of reserves and incs and decs could jeopardize real-time operations. More subtle are the downstream effects of incomplete forward markets. Forward trades on an hourly basis do not fix output rates over the hour, so rapidly changing demand within an hour, such as the initial morning ramp up, are often met with heavy doses of regulation or other reserves. More generally, the few categories of reserves for which day-ahead markets are conducted limit the SO's flexibility; e.g., when these markets omit decremental reserves. The SO's anxiety is part of the motive for purchasing more reserves than centralized systems do, but another part is the greater volatility of decentralized markets. In a fully decentralized system the SO does not control or direct dispatch except via the inc/dec options it invokes, so suppliers can deviate from day-ahead schedules, leaving the SO responsible for balancing the system based on their actual outputs (which are metered only *ex post*). The potential deviations can be large if, say, suppliers bypass the forward markets because they expect higher spot prices, or demanders because they expect lower spot prices. (Single-settlement systems are immune to such arbitrage, but they are vulnerable to withholding of supplies to increase the spot price on which day-ahead transactions are settled.) Because arbitrage among forward and spot markets is necessary to keep their prices linked, ideally as a martingale, designers avoid penalizing large deviations except when they cause market failures, as in the case of onerous default prices imposed in California.

From this overview of real-time operations and spot markets it appears that decentralized systems are inferior. The SO in a centralized regime can re-optimize the entire system every few minutes to re-dispatch all feasible resources, whereas in a decentralized regime the SO has weaker control of a more volatile system – and both the weakness and the volatility stem from imperfections in the market structure. No decentralized system shows signs of lesser reliability, but there is evidence of higher costs for reserves. One interpretation is that, contrary to appearances, the reserve markets differ in timing but not in substance. Centralized systems require participants to maintain (or acquire in secondary markets) sufficient installed and operable capacity and to bid all operable capacity into the market, enabling the SO to allocate any portion to reserves or the spot market, whereas in decentralized systems the SO conducts a daily auction to procure sufficient reserves – the end result could be the same, but evidently the decentralized implementations do not yet match the performance of the centralized procedures.

What then are the purported advantages of decentralization? The answer to this question requires study of the forward markets. We continue in reverse order to examine the markets for reserves, transmission, and energy.

3.2 Forward Markets for Reserves

Centrally optimized systems use suppliers' initial energy bids to assign some to reserve status, typically compensating those curtailed for spinning reserve the amount of their profits foregone in the energy market and paying the bids of extra-marginal units. Even so, all operating units are subject to re-dispatch in real time.

In contrast, participants in decentralized systems can either self-provide the required percentage of reserves or buy it from the SO, who procures sufficient amounts of each category in a series of auctions conducted day-ahead, and additional resources contracted

long-term. The design of reserve markets has had a tortuous history that stems from three complications.

The first remains from the era of vertically integrated utilities with universal service obligations and simple tariffs. Most systems make little use of reserve options on the demand side, such as contracts for service that can be curtailed by the SO. Competition at the retail level stimulates demand-side participation in the reserve markets, but progress is slow, due partly to the initial expense of installing smart meters.

The second is that the categories of supply reserves are substitutes in a quality hierarchy derived from response times. The faster response time of regulation implies that it can substitute for spinning reserve (but not reversely), and similarly spinning reserve can substitute for non-spinning, *et cetera*. This implies that all reserve markets must be cleared simultaneously, with the result that prices decline as response times increase. In California, the initial implementation provided a separate auction for each category and the SO's demand in each was specified inelastically. Instances of prices increasing with response times revealed the problem, but not before prices for some low quality reserves were a thousand times normal levels. Subsequent efforts to design procedures and software to clear the four main reserve markets simultaneously while taking account of the implicit uni-directional substitutability is a lesson in the practical difficulties of implementing markets for multiple goods that are substitutes or complements, even when the theory is clear and simple.

The third complication is that a reserve bid has at least two parts or dimensions, and so do settlements. One part is the price offered for capacity availability and the other is the price offered for energy generated when the SO invokes its option. The theory of multi-dimensional auctions is more complicated, and judging from occasional disasters, so is practical implementation. The usual fallacy is to combine the two parts by using a scoring rule and accepting those bids with the lowest scores until demand is filled. For example, the score could be the capacity bid plus the product of the energy bid and the expected quantity of energy generated. If this expected quantity by which the energy bid is weighted is not optimally determined as a complicated function of all bids – usually it is just a constant based on the SO's prediction of average energy requirements – then a flood of unfortunate efficiency and incentive effects ensue. The first effect is that the real-time energy payments do not conform to the merit order in which options must be exercised to preserve efficiency. Another effect on efficiency is that the scoring rule can attract low-cost supplies that optimally should be sold in the day-ahead energy market – this effect occurs whenever the SO seeks to minimize the cost of its purchases rather than to maximize the gains from trade in all markets combined. The incentive effects can be extreme. Each bidder recognizes that his actual chances and duration of energy generation depend on his energy bid rather than the SO's predicted average, so he sees a tradeoff between the capacity and energy parts of his bid that encourages distorted reporting of costs. In the worst case, he thinks the SO's prediction is wrong, say too high, in which case the optimal bid inflates the capacity part and deflates the energy part to zero (or negative in the notorious case of the 1993 BRPU auctions in California).

Fortunately, a two-dimensional reserve auction can be reduced to a one-dimensional auction by the simple device of treating the energy bid as a reservation price and settling accounts for actual energy generation at the spot price. That is, the scoring rule for the auction of capacity availability comprises merely the capacity bid, with zero weight given

to the energy bid. The energy bid is interpreted as the spot price below which the supplier prefers not to be called for real-time generation, so in effect the energy bid becomes the price of its inc or dec in the merit order. The optimality of this auction design depends, however, on separate markets for incs and decs.⁸

Even though the complications described above have solutions, it is important to emphasize that reserve markets are widely viewed as the weakest link in decentralized designs. To some extent this is inevitable when few demand-side options are available, forcing the SO to juggle supplies in real time to meet demands that include significant stochastic and cyclical variations. Providing the SO with ample flexibility seems to require many markets – at least 4 categories of reserves that are partial substitutes, one or two of which should include decs as well as incs, and one adapted to load following. Perhaps better would be a unified market differentiated by a quality dimension (response time) whose remuneration is determined as the SO's opportunity cost of substituting the bid from the next slower unit. The ultimate solution, however, is to enrich the reserve options obtained from the demand side.

3.3 Forward Markets for Transmission

The design principles for transmission markets are broadly similar in electricity, gas, telecommunications, rail and other transfer networks affected by congestion. The distinctive features of electricity are that a point-to-point injection and withdrawal of energy dissipates a portion as heat, which I will ignore henceforth, and that the flow of electrons among alternative paths obeys Kirchhoff's Laws, so it is largely uncontrollable in systems with alternating current. When the net demand for a transmission link exceeds its safe transfer capacity it is said to be congested; remedies include both demand reduction and provision of counterflows to reduce the net flow. An uncongested transmission system resembles a pool to which one can add or subtract water, so in effect it unites all suppliers and demanders in a single marketplace.⁹

Many systems worldwide are designed to eliminate virtually all congestion on the grounds that the transmission system is a necessary part of the infrastructure for an efficient industry. It is a public good due to technical externalities, and also due to pecuniary externalities since contestability depends heavily on sufficient transmission capacity. When this is accomplished by building ample capacity, an access fee is charged to recover construction and maintenance costs. In those decentralized systems with limited transmission capacity, the SO alleviates congestion by intervening in the energy markets to purchase counterflows on congested links; again an access fee and perhaps a uniform injection charge recover costs. Centralized systems reduce flows or produce counterflows by directing various generators to contract or expand energy output, providing compensation based on bids.

The alternative approach uses market processes to establish energy prices that are differentiated by location and therefore induce the required counterflows. Centralized

⁸ H. Chao and R. Wilson (1999), "Incentive-Compatible Evaluation and Settlement Rules: Multi-Dimensional Auctions for Procurement of Ancillary Services in Power Markets."

⁹ Gas transmission differs slightly. Apart from pressure maintenance, it is a displacement system in which the gas in the pipe, called linepack, is merely displaced by an injection at one point and withdrawal of an equal quantity at another point. Some reserve can be obtained by varying the pressure in the pipe. Longer term reserves are provided by underground storage.

systems obtain the energy price at a node as the shadow price on an injection there, or equivalently, by constructing it as the sum of the system price for energy plus an injection charge: the injection charge is derived from the shadow prices on the capacities of all transmission links by using Kirchhoff's Laws to predict the distribution of flows on links produced by an injection. In a large system like PJM, fully differentiated pricing requires setting prices at thousands of nodes, or on thousands of links, but this complexity is often reduced by setting nodal prices only at major hubs, or uniformly across large zones as in California.

Decentralized markets rely on incs and decs to alleviate congestion. To simplify, suppose there is congestion on lines from an exporting zone to an importing zone. That is, clearing the energy market would result in a single price (the "uncongested" price) and a flow exceeding the transmission capacity. The remedy in Scandinavia's NordPool is to raise the price in the importing zone and reduce the price in the exporting zone until the net flow matches the capacity; the difference between these two zonal energy prices is then the usage fee charged for flows from the exporting zone to the importing zone – and equal credit is given for counterflows. In effect, NordPool uses the inframarginal bids in the supply and demand functions submitted in each zone as offers to increment or decrement energy output. This illustrates the general principles that transmission demands are derived from energy demands and supplies, and like reserves, congestion is managed by amending the forward market for energy, but unlike the simultaneous optimization of all three aspects attempted by centralized systems, decentralized markets for energy, transmission, and reserves operate in sequence.

California's transmission market is similar, but in keeping with its pervasive theme of voluntary participation, it allows bidders to submit incs and decs to the transmission market that differ from their bids in the previous energy market. Sometimes the SO receives insufficient offers to alleviate congestion and the market fails to clear, in which case a default usage fee is imposed. The default fee is partly punitive, but also it is intended to cover the SO's expected costs of fixing the problem in real time using incs and decs offered in the spot market or by invoking reserves. The occasional collapse of purely voluntary markets is another example of the seeming fragility of decentralized designs.

Those systems that set usage fees only between large zones reflect compromises among competing objectives. Usage fees based on markets for alleviating congestion are universally recognized as the efficient design based on theoretical considerations. Arguing against this are practical motives. One motive is to minimize the intrusions of the SO into the forward energy markets, due to apprehensions about inherent monopoly power derived from its exclusive control of the transmission system. This stems from the practical consideration that nodal pricing is apparently feasible only within a comprehensive optimization of energy and transmission conducted by the SO. A related practical matter is that efficiency gains from elaborate nodal pricing in forward markets are likely small given the subsequent repetition of congestion management in real time, and the usual pattern that only a few main inter-ties are congested. Another motive is to maximize the competitiveness of the forward energy markets by creating a common marketplace, which zonal pricing does by ignoring congestion within each zone for the purposes of forward markets. Zonal pricing also serves, of course, as a mutual insurance scheme among participants within each zone.

This compromise creates adverse incentive effects, however. Zonal pricing in a highly decentralized system like California enables strategies like the following. A supplier who anticipates intrazonal congestion affecting his injection node can sell a quantity $2Q$ in the day-ahead energy market at its clearing price P when he knows that in real time the SO will be forced to invoke the dec he offers for the quantity Q at the spot price p^* , which is typically lower than P when decs are invoked, or at his bid price p , which is even lower when the dec is invoked out of merit order. The net result is that the supplier collects a profit $[P-p^*]Q$ or more on the extra quantity Q that he knew initially he would not produce.

These incentive problems are compounded when establishing conditions for entry. Nodal prices need not include all external effects on congestion in the network caused by new plants, but at least they recognize the local effects. Zonal pricing ignores local effects, and via its access charge the SO spreads among all participants its cost of alleviating intrazonal congestion in real time, so in attempting to force entrants to pay the costs they impose on others, incumbents like to devise schemes that raise barriers to entry. The gaming strategy outlined above provides a clue about a measure that might help: require that for some period the entrant pays the spot price for a dec even if it is invoked out of merit order. This eliminates entry justified only by exploitation of gaming opportunities of the sort described above. Some measure along these lines is necessary since otherwise an entrant prefers to build in the most congested location, precisely the opposite of what is required for efficiency. To deter abuses the SO could threaten to define a new zone around the entrant's plant, but this amounts to nodal pricing.

The source of these problems is incompleteness of the transmission market. Intrazonal transmission is not priced explicitly in either the day-ahead or spot market – even when it is known *ex ante* to be a scarce resource requiring the SO to exercise decs out of merit order at whatever price a supplier bids. Scandinavia's NordPool addresses this deficiency by using zonal definitions that can be changed daily to conform to the pattern of congestion, but the rigid zones in California preclude this method. A transition to nodal pricing, at least at hubs or in small zones, seems inevitable.

A persistent tension in transmission markets stems from participants' insistence on financial hedges against usage fees, and even firm rights to physical access like those sold by gas pipeline companies. In fact, the U.S. regulatory agency requires each SO to provide "price certainty" for transmission. This requirement is satisfied when the SO offers long-term transmission "rights" in an auction, and facilitates trades in secondary markets. The source of the demand for hedges and rights may be due to genuine risk aversion, but more likely it reflects marketing advantages obtained by brokers who bundle transmission rights with energy transactions in long-term bilateral contracts.

A financial right entitles a buyer to a continual refund of the usage fee whether or not he transmits energy, and when the right includes a scheduling priority, physical access is virtually assured. In centralized systems like PJM, a financial right specifies an injection point and a withdrawal point, which is apparently necessary to conform to optimization procedures in which the bids are interpreted as point-to-point balanced injections and withdrawals for purposes of simulating operations to derive nodal prices, and thereby deriving the auction price of a right as the difference between the nodal prices. This point-to-point definition limits resale and stifles secondary markets so provision is made for periodic reconfiguring of the collection of point-to-point rights.

In decentralized systems like California, each right pertains to the interface between two zones and includes both a financial hedge and scheduling priority, which together amount to a lease. A peculiar aspect is that leasing 100% of interzonal transmission this way amounts to privatization, and it implies complete reliance on secondary markets to allocate interzonal transmission because using incs and decs to alleviate congestion becomes ineffective due to the rights' absolute priority for scheduling. Recent research predicts that hedges against transmission fees magnify the market power of suppliers in import zones.¹⁰

A general issue that pervades the economics of transmission markets is the effect of market organization on allocative efficiency. As mentioned, the demand for transmission derives from energy transactions. If the energy market is conducted as a call auction then the demand value of transmission is expressed accurately in terms of the gains from trade that transmission enables, as in NordPool's method for instance. With bilateral trading, however, random matching of buyers and sellers creates for each pair a gain from trade (= their joint demand value for transmission) that alters the derived aggregate demand curve for transmission. For example, using incs and decs to alleviate congestion need not be efficient when the pairs whose trades are curtailed are not the ones with the smallest gains from trade. The practical importance of this feature need not be important if brokers remedy the problem, but otherwise it indicates a role for central exchanges with market-clearing prices to handle some percentage of trade. In many countries trades are mostly bilateral but still the day-ahead exchange handles 10 to 20% of the trading volume, which is usually enough to ensure efficient allocation of transmission.

3.4 Forward Markets for Energy

The variety of designs used in energy markets is remarkable. At this early stage it is unclear whether variety offers permanent advantages or the industry will eventually converge to one or a few designs. Evolution, but not necessarily progress, is evident in Britain's switch from a central exchange to decentralized markets for bilateral contracts. I describe some general aspects and then examine two dimensions along which designs differ.

The SO's time frame for operational control spans an hour or two, and day-ahead planning is sufficient to purchase reserves, schedule voltage support, etc. In fact, Britain's new system provides the SO with just 4 hours advance notice of energy transactions. Such short horizons are possible because in principle an SO accepts only balanced schedules in which energy injections equal withdrawals, so it is only in real-time operations that an SO must cope with imbalances.¹¹ In most systems, however, day-ahead notice is required to provide ample time to alleviate anticipated congestion on major transmission lines. California and PJM, for instance, use day-ahead markets to balance transmission on major lines so that real-time operations handle smaller local deviations.

¹⁰ P. Joskow and J. Tirole (1999), "Transmission Rights and Market Power on Electric Power Networks I: Financial Rights, II: Physical Rights."

¹¹ Violations of this principle exacerbate problems in real time operations. Examples are failures to account for losses or for energy from units providing voltage support or reactive energy.

This sequence of day-ahead, then real-time, operations for the SO meshes with longer time frames in the energy markets.¹² For thermal generators, the basic scheduling decisions are unit commitments (startup, ramping, running rates) made daily, so in systems with substantial thermal capacity, prices in day-ahead forward markets are basic to productive efficiency. Real-time energy demand can typically be predicted day-ahead within 3% for each hour, so day-ahead scheduling largely suffices. Longer commitments are made via bilateral contracts, some of which are physical contracts for actual production and delivery, and others, financial hedges. Within the operating day, deviations from initial schedules are common, due mainly to demand variations addressed via the spot market and by invoking options on reserves. Mature systems show a pattern of up to 80% contracted long term, 20% day-ahead, and less than 10% spot, although much of the supply contracted long-term actually passes through the day-ahead market. Contracts are often specified as contracts for differences (CFDs) in which the parties mutually insure each other against the difference between their contracted price and the market price.¹³

Because centralized systems consolidate all energy markets, the basic structure of the forward markets is better described in terms of a decentralized system, using California as the archetype. I divide the topics between organizational forms and trading arrangements.

Organizational Forms

The two main organizational forms are adapted to the contracts traded. In contracts for physical delivery, the counterparty is either another market participant or the market manager.

- Among those contracts between participants, essentially all are bilateral because multilateral contracts are practically infeasible. The market manager (if any) in such cases functions essentially as a broker. Some bilateral markets are merely electronic bulletin boards on which bids and offers are posted, and others offer standard contracts; e.g., one is a 5×16 contract for delivery over five weekdays in the sixteen peak hours. Auxiliary terms and conditions, and bundled hedges against transmission and reserve prices, simulate some aspects of markets conducted by dealers, but dealer markets for pure energy are precluded by the non-storability of electricity.
- Those contracts in which the market manager is the counterparty are conducted as exchanges in which the manager balances aggregate demand and supply, and uses receipts from demanders to pay suppliers, net of losses.

Both brokers and exchanges charge transaction fees. The contracts are termed physical because delivery is expected, but actually all forward transactions are inherently financial since commitments can be reversed by purchases or sales in the spot market. In both forms the typical pattern is for a participant to contract forward based on expectations but then to adjust based on contingencies arising the next day. An SO's procedural rules include specific assurances that balanced energy schedules submitted directly (from a few

¹² The gas industry is similar. An SO or a pipeline company does day-ahead and intra-day scheduling while the commodity markets use long-term contracts, a monthly planning horizon, and daily scheduling.

¹³ Similarly declining percentages can be seen in fuel markets such as gas and other commodity markets, including even metals, but there is an increasing tendency towards more short-term trading as electronic communication and controls improve to allow more demand-side responsiveness to spot prices.

large participants allowed direct access to the SO), from brokers of bilateral contracts, or from exchanges are all treated comparably, so in principle there is no bias in scheduling transmission or reserves.

The division of the market between long-term contracting directly or through brokers, and short-term (day-ahead or day-of) through power exchanges is partly an artifact of the institutional arrangements. With a few exceptions, exchanges are established at public expense as non-profit public-benefit corporations by legislation that confines their scope to short-term markets (although a few also conduct supplementary markets for longer-term CFDs hedged against the exchange price). Their purpose is to ensure a transparent and liquid forward market whose prices can be used as benchmarks less volatile than spot market prices. Markets for purely financial instruments such as futures contracts expand the influence of exchanges because they are used mainly as hedges against the exchange price and they are based on the exchange's delivery points and conditions.

However, it is Britain that established one of the first day-ahead exchanges in 1989 and now intends to abolish it, relying entirely on bilateral transactions in private markets. Even though exchanges have successful records from Scandinavia to Australia, the necessity and viability of exchanges remain doubtful. California requires its power exchange to compete with bilateral markets (and also a private exchange) but others provide the exchange with a monopoly on short-term trades and some require bilateral contracts to pass through the exchange. If exchanges whither then their public good – a liquid and transparent market – is likely to vanish since brokered markets for bilateral contracts are intensely secretive. Efficiency could be affected because monitoring and controlling market power become difficult, and ultimately the market power of dominant brokers must be addressed.

Trading Arrangements

Few details are known about how bilateral contracts are privately negotiated or facilitated by brokers. In the U.S. and Canada, several major suppliers engage in active marketing, employing traders who solicit deals and exploit arbitrage opportunities. Markets conducted via bulletin boards for posting bids and offers for standard contracts use simple trading arrangements; similarly markets for CFDs and swaps are conducted by telephone. The chief complication in these markets is counterparty risk, the chance that the other party to the transaction will default – a notorious episode in 1998 convulsed the midwestern U.S. market due to domino effects on other parties, including bankruptcies. An evident advantage of public exchanges is reduced counterparty risk.

In contrast, exchanges rely on sophisticated trading arrangements. Their authority to experiment is invariably restricted; for instance, an innovation like a Vickrey auction is precluded by prohibitions against price discrimination and a mandate to clear each hourly market independently at a uniform clearing price. But within these restrictions they have broad authority to promote efficiency. For example, the bid format is fairly rich, enabling each participant to submit a supply or demand function to each hourly market. These bids, moreover, are for energy only so that afterwards a supplier can conduct its own optimization of unit commitments and operating schedules. This requires internalization of startup costs, ramping constraints, and other considerations but on the other hand, given the total energy sold in the market, it encourages productive efficiency using the supplier's private information about its costs. The Mercado in Spain offers another

example: it allows withdrawal of tentatively accepted bids that do not meet the minimum revenue required to justify startup. Proposed designs elsewhere allow a bid format that enables a supplier to specify a minimum duration and a minimum output rate for each thermal generator. Another (for which I designed activity rules) enables bidders to take account of intertemporal considerations: it uses an iterative auction so that participants can revise their bids in response to the observed pattern of prices over the 24 hourly markets for next-day delivery.¹⁴

The deficiencies of exchange procedures are obvious to economists. The bid format and market clearing procedures take little or no account of intertemporal factors, and no contingent contracts are traded. Settling trades at a uniform clearing price encourages withholding of supplies by firms with market power, and excludes a Vickrey design and most other means of strengthening incentives. The clearing price is only that, it does not necessarily represent accurately the actual opportunity cost derived from shadow prices in a full system optimization.

The strengths are less obvious but significant. Prices are more reflective of actual costs because suppliers schedule their plants. Settling the forward and spot markets at their own prices suppresses gaming to affect the spot price and optimally penalizes deviations by using the spot price. It also promotes arbitrage between the forward and spot markets, and more correctly rewards flexible resources such as peaking generators. Active bidding by demanders is encouraged; in contrast, those centralized systems operating presently have no significant demand-side bidding. A feature valued by participants is that prices are derived transparently from bids with no opaque model and arcane algorithm intervening to compute shadow prices.¹⁵

Theorists have worried most about the consequences of sequential markets. The sequence of energy markets (day-ahead, day-of, spot), and the parallel sequence of transmission and reserve markets, depend on rational expectations that could go awry. The extreme case occurs in California where, unlike NordPool, a supplier's offers in the day-ahead market for energy cannot be conditioned on the usage fees charged for interzonal transmission, nor on prices in the reserve markets.

4. Mitigation of Market Power

In most countries, the process of market liberalization begins by privatizing state enterprises; in others, vertically integrated utility companies are divided into units that are functionally specialized in generation, transmission, distribution, or retail sales. Continued regulation of the "wires" businesses of transmission and distribution is standard but system operations in some jurisdictions are assigned to a non-profit enterprise called the ISO or IMO (independent system or market operator) charged with ensuring that the transmission system provides infrastructure for efficient markets. Anxiety that the transmission owner (Transco) would exert monopoly power or favor affiliates, even when regulated, may stem from ample evidence about the practices of gas

¹⁴ R. Wilson (1999), "Activity Rules for an Iterative Double Auction," chapter 10 in K. Chatterjee and W. Samuelson (eds.), *Business Applications of Game Theory*, Kluwer Academic Press.

¹⁵ The aversion to optimization models and algorithms is surprisingly deep in California, where they were called Gosplan after the USSR's central planning agency. In Britain the performance of the software implementing the GOAL algorithm was erratic, but it could not be modified because so many contracts were tied to the prices it computed.

pipeline companies in the U.S.¹⁶ Of course one must be equally anxious about whether the governance structure of an ISO is sufficient to ensure efficient operations without favoritism. No proposal for management of the ISO as a franchise is sufficiently developed to ensure perfect incentives for efficiency. In either case, the Transco or ISO inherits the software and technical personnel familiar with system operations.

With regulation pervasive in other aspects, and some confidence that retail sales will be competitive, concerns about market power focus on generation.¹⁷ Few countries are eager to break apart the generation operations of a well-functioning state enterprise or legal monopoly, especially one owning major assets such as nuclear plants and hydro reservoirs, so invariably much attention is given to clever ways of mitigating market power. Even those bold enough to require divestiture of generation assets have avoided strict requirements sufficient to ensure vigorous competition, Britain being the prime example. The boldest have been states in the U.S., but even so they have allowed full-fleet sales in which a large segment of the generation capacity is sold to a single buyer – who in many cases is just the unregulated affiliate of a utility from another state engaged in a game of musical chairs.

Apart from divestiture, there are several direct ways of mitigating market power.

- (a) One is to require that reliability-must-run (RMR) plants, which have monopoly power because they are occasionally needed for local voltage support, must operate under long-term contracts with remuneration based on audited costs. When reserve margins are thin, similar provisions could be applied to plants with unique capabilities to meet peak loads. This seemingly simple approach is difficult in practice: over a third of capacity in California was assigned initially to RMR contracts, and various strategic manipulations were persistent problems, in part because the energy produced was not matched by demand in the forward market so a surplus of energy supplies spilled into the spot market.
- (b) Another common way is to require firms with market power to be heavily hedged by long-term forward contracts for delivered energy at fixed prices, sometimes called legislated hedges or contract cover. In Britain the hedges purportedly worked well to sustain incentives for output until the initial contracts expired. The same might be true of Alberta but the percent hedged was so high that price variation was damped, and entrants were excluded by hedges contracted between each company's generation and distribution subsidiaries – in fact, hedging was so pervasive that prices in the spot market served primarily as transfer prices between subsidiaries.
- (c) There are several ways to simulate the effects of hedges. The simplest requires dominant firms to auction some percentage of their output, usually in the form of long-term contracts, as in Alberta's revised design.¹⁸ An important proviso is that

¹⁶ In the US, pipelines' monopolistic practices include discrimination in terms and prices (except firm service at the maximum price allowed by regulators) and withholding of capacity. An instructive contrast is between pipelines' practice of charging for "parking and lending" services and Australia's VicGas system in which an end-of-day balancing market provides equivalent services. The US system allows a pipeline to lend one shipper's excess to another who is deficient and charge them both.

¹⁷ Few worry about low prices due to monopsony on the demand side of wholesale markets, partly because retail demand elasticity is low, but this sanguine attitude erodes as reserve margins shrink due to insufficient entry. I ignore monopsony power here.

¹⁸ Proposed regulation of gas pipelines in the US requires daily auctions of all residual capacity. This focus on the spot market is more viable when capacity is defined by simple measures like throughput capacity.

buyers have full latitude to resell purchased supplies in competition with the firm, and further, the contract terms must discourage sellers from using operating and maintenance decisions to disadvantage buyers, and preclude repurchase agreements.

Another way is to link ownership of supply capacity with enough demand-side obligations to make the firm a net buyer, who therefore may prefer low prices. This structural solution was implicit in early configurations of the California and New Zealand industries, and in countries like Norway where local distribution companies own substantial capacity.

In developing countries, liberalization often applies initially only to wholesale markets and distribution companies retain monopoly franchises in retail markets; in such cases it might suffice to require a dominant supplier to auction entitlements to distribution companies in the form of contracts for “virtual capacity” in which the firm manages assets and operations but passes variable costs through to the distribution company.

The relevance of these measures depends on how contestable the market is. No action is needed if a flood of imports from contiguous regions would erode the market power of a dominant incumbent. But this conclusion depends on the availability of sufficient transmission capacity, and this in turn depends on the governance of the entity that decides on expansions of transmission capacity. An incumbent can easily argue that expanding capacity is unjustified if it will be idle, whereas in fact it is idle capacity that importers require to offer supplies in competition with the incumbent. The incumbent should not be able to veto expansions proposed by representatives from the demand side.

Governance arrangements are potentially hazardous to entrants. Suppliers on the governing board of an ISO can argue for technical requirements or compensation that amount to barriers to entry. These impediments are most likely in systems organized as legal cartels, as in New Zealand. I mentioned earlier how zonal pricing enables incumbents to argue that entrants should pay for additional congestion they cause, because without nodal pricing there is no direct measure, and further there is a perverse incentive to enter at a congested location to garner payments for decs invoked out of merit order in the spot market (recall that a partial remedy is to allow entrants only the spot price for decs). The magnitude of the administrative hurdles are amazing: at last count an entrant into New England must win approvals from twelve technical panels of the regional power pool association.

A peculiar feature of those systems derived from power pools are payments to attract further capacity expansion and to retain obsolete plants in working order. Requirements for installed and operable capacity also induce capacity payments when these obligations are tradable in auxiliary markets. A typical example was Britain’s capacity payment based on the product of an estimated probability of insufficient capacity, and a value of unserved demand that was set administratively. (As a cynic might expect, the estimated frequency of outages was ten times the actual frequency.) Theory establishes that indeed capacity payments are optimal, and should equal the capacity cost of the most efficient peaking generator, perhaps a combustion turbine. This payment seems necessary because such a generator is idle most of the year. In fact, however, this theory is an obsolete remnant of an era in which demand side responses were ignored. Demanders who accept contracts allowing loads to be curtailed or interrupted are usually the most efficient

substitute for peaking capacity, and their capacity costs are nil so no capacity payment is required.

This reflects a wider aspect, which is that incumbent suppliers have no incentives to encourage demand-side bidding, and actually prefer that the predicted load from demanders is represented inelastically in the forward markets. Opportunities to make markets more contestable by encouraging demand-side bidding are easily ignored even when they offer the brightest prospects for long-run efficiency of the industry.

Against these considerations must be counted the positive effects of relational contracting in centralized systems. New Zealand takes the most aggressive stance: its Market Surveillance Committee has broad powers to implement efficiency-improving changes to the market rules and to sanction abuses of market power. In contrast, similarly named committees in decentralized systems usually have authority only to monitor performance and to address occasional reports to regulatory agencies – and in Alberta the committee resigned in frustration.

5. A Concluding Perspective

The peculiar aspect of centralized designs is diligent effort to optimize and coordinate operations with cursory attention to incentives. This inheritance from the era of vertically integrated utilities is insufficient when some participants have market power sustained by barriers to entry and imports. Skimping on incentives is unnecessary because, for example, the settlement rules of a Vickrey auction or its Bayesian counterpart could encourage accurate reporting of costs and discourage implicit collusion. Such rules entail price discrimination, and in fact pay tribute to powerful players, but they enable optimization procedures to attain the intended objective of productive efficiency. Given the complexity of existing optimizations, the usual complaint that incentive-compatible pricing is too complicated is not salient. But aversion to price differentiation runs deep.¹⁹

Decentralized designs suggest a different path to improving performance. The importance of incentives in getting prices right is recognized explicitly. Rather than shadow prices derived from direct revelation games, they use clearing prices in auction markets so that strategic behavior focuses on bids that affect prices directly and transparently (absent collusion in repeated auctions). The role of the SO is minimized and suppliers self-schedule plants to meet their sales of energy. This division emphasizes the difference between instantaneous control of the spatially differentiated transmission system, and scheduling plants efficiently to satisfy intertemporal constraints.

The inevitable deficiencies are tenuous control of transmission and reserves, and incomplete and imperfectly coordinated markets. Most problems can be traced to weak links between forward and spot markets, among the separate markets for energy, transmission, and reserves, and within each of these, among its hourly sub-markets. But incompleteness stems from static procedures as well as simplified bid formats and market-clearing rules. It is well established, for instance, that more frequent trading opportunities enhance the completeness of a market; thus, a rich sequence of forward markets for simple contracts approximates a single complete market. The California Power Exchange, for instance, conducts long-forward, day-ahead, and multiple day-of

¹⁹ Even theorists are troubled by examples of Vickrey auctions in which, due to mixtures of substitutes and complements, two suppliers are paid more than another who bid less to supply their combined output. For practitioners the aversion is to the fact that the settlement rules of a Vickrey auction reward market power.

markets for energy, complemented by the SO's real-time spot market and parallel forward markets for bilateral contracts. It seems likely that this sequence accomplishes most of what a market for more complicated contracts might achieve.

Similarly, the more iterative is each auction, the less a participant needs complete markets and the more factors and contingencies he can internalize when optimizing his own operations. That is, better price discovery brings better performance. If price discovery were perfect, then the existing set of forward and spot markets used in wholesale markets for energy, transmission, and reserves would suffice. This a lesson implicit in standard theories of Walrasian markets that invoke the magic of an invisible hand to establish market-clearing prices.

Only recently have procedural devices been developed to improve price discovery. The prime example is the activity rule used in iterative auctions of spectrum licenses. This rule forces each bidder to reveal its demands early by restricting subsequent opportunities to a subset of those in the current iteration, based on those it actually bids on. My design of an iterative energy auction for California introduced a similar "use it or lose it" option: a bid rejected in one iteration and not revised in the next to better the previous clearing price cannot be so revised later.²⁰ These devices are crude compared to innovations that progress in auction theory and market design will provide. I see them as indications that procedural rules for dynamic auctions can improve price discovery, and in turn, good price discovery enables a small set of simple markets to approximate complete markets.

The implication is that economic theory can contribute usefully to market design by telling practitioners more about how structural design features affect the dynamics of imperfectly competitive markets. The features of the electricity industry suggest studying a sequence of incomplete forward markets, each of which might be an iterative auction to promote price discovery. It seems plausible that if there is little market power then decentralized designs are preferable when participants have significant private information about optimizing their own operations. The deeper issue is how the choice between centralized and decentralized designs, or hybrids, is affected by substantial problems of market power.

Each liberalized market in an infrastructure industry offers another challenge, with more to come. Each industry exhibits tension between a centralized mechanism that gets the prices right ignoring incentives, and decentralized markets that enhance competition within limitations imposed by incomplete contracts and imperfect price discovery. Given the trend to decentralization, I hope economic theory can keep up with the rapid progress that practitioners are making in improving price formation processes in these industries.

²⁰ An intriguing aspect of an activity rule is its effect on the process of competition. The rule above encourages a supplier who can profitably reduce its bid below the tentative market clearing price to do so, lest later opportunities to do so are lost. Reducing the offered price of such an extramarginal bid then displaces some inframarginal bid that was tentatively accepted, and which must then also be reduced if the opportunity is not to be lost. This induces a struggle among the marginal bidders that drives the clearing price down in successive iterations. Competition among marginal bidders is the sort envisioned by Marshall, in contrast to the static competition via submitted supply functions described by Walrasian models and implemented in centralized systems.