Econometrica, Vol. 70, No. 4 (July, 2002), 1299–1340

ARCHITECTURE OF POWER MARKETS[1]

BY ROBERT WILSON[2]

Liberalization of infrastructure industries presents classic economic issues about how organization and procedure affect market performance. These issues are examined in wholesale power markets. The perspective from game theory complements standard economic theory to examine effects on efficiency and incentives.

KEYWORDS: Market design, liberalization, regulation, electricity.

1. INTRODUCTION

A PROCESS HAS BEEN UNDERWAY worldwide for three decades to privatize state enterprises and liberalize markets for the services of infrastructure industries, including water, communications, electricity, fuels such as gas, and transport by airlines, trucks, and railroads. This process is usually viewed as replacing tight regulation of vertically integrated monopolies with light regulation of functionally specialized firms and supervision of competitive markets. The shift was justified by changes in technology, such as diminished economies of scale exemplified in the electricity industry by smaller efficient plants. In airlines and trucking, contestability was viewed sufficient to limit market power, and in telecommunications contestability was enforced by requiring incumbents to offer wholesale tariffs to resellers, and in some cases access by competing carriers. Some countries simply established the transport network—such as power, gas, or rail lines—as a common carrier separate from the commodity or service industry. Another view emphasizes the role of unbundling to expose cross-subsidies and to improve efficiency via better pricing and stronger incentives for product variety. A prevalent view in developing countries sees privatization and liberalization as necessary to overcome organizational inertia and to attract new investment.

This essay examines liberalization to find lessons relevant to economic theories of market microstructure. The normative tone reflects the increased role of

[1] 1999 Presidential Address to the Econometric Society presented at the North American, Far Eastern, Australasian, and European regional meetings. Wilson (2001b) provides an expanded version of Section 3 and includes investment issues not addressed here. This revision includes material in Section 4 on later events in California.

[2] Research support was provided by the Electric Power Research Institute and National Science Foundation Grant SBR9511209. I am grateful for joint work on these topics with Hung-po Chao and Shmuel Oren, to several colleagues including Peter Cramton, Preston McAfee, John McMillan, Paul Milgrom, Charles Plott, Alvin Roth, and Frank Wolak for shared interests in these topics, and for long collaborations to Srihari Govindan, Faruk Gül, David Kreps, and John Roberts. Paul Joskow and Jean Tirole provided valuable comments on the 1999 version, and Stephen Peck, John Roberts, and a referee on a draft of this version.

economics as an engineering discipline capable of providing guidance on details of market design. This role grew as game theory and derivative theories of incentives and information expanded economists' tools to include methodologies for predicting how procedural aspects influence participants' strategies and affect overall performance. Part of this toolkit pertains to the standard concerns of economic policy such as productive and allocative efficiency and mitigation of market power; 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—*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 new wholesale markets conducted as auctions in the electricity industry. These provide a rich context for issues of market design, and further, they illustrate the principle that designs are tightly constrained by technology. No two designs among the liberalized power markets are the same, so in effect an enormous experiment is underway and one can learn from comparative studies. I use the Northeast and California systems in the U.S. as archetypes.

Within this narrow focus, I discuss three issues: the extent of reliance on markets, detailed design of forward and spot markets, and allocation of risk. The Appendix sketches methods of limiting market power. First I embed these issues within a larger context in the economic theory of markets. For readers seeking further details on the structure of wholesale and retail markets for electricity, I suggest Stoft (2002).

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 power is a flow (or field) of energy that cannot be monitored perfectly, and storing potential energy is expensive; many of the unique features of electricity markets stem from these two features. Also, flows on transmission lines are constrained continuously by operational limits and environmental factors, and ramping rates of generators are limited. But the primary cause is variable demand that presently is not matched with flexible spot pricing at the retail level, except for large industrial customers equipped with real-time meters, and in any case the short-run elasticity of demand is notoriously small. Devices for continual metering and control are feasible (though expensive), and innovative tariffs and service plans have been introduced on small scales, but implementation was slow before liberalization, and paradoxically, often retreated after liberalization (Section 4 discusses some of the adverse consequences of liberalizing wholesale markets before developing price responsiveness in the retail sector).

In the long run, imperfect competition in power markets stems from the same factors as in other industries, such as economies of scale and other entry barriers, and oligopolistic ownership. Competition is imperfect in intermediate time frames because production is capital intensive and construction delays are long compared to variations in supply and demand conditions. On short time scales, prices are inherently volatile and competition is often imperfect because of technical rigidities on the supply side, and inelastic demand, sketched below.

Power transfers are complicated by the difficulty of directing flows in transmission systems with alternating current. Between points of injection and extraction, flow occurs on each possible path in inverse proportion to impedance (the new technique of "phase shifting" allows some directional control by altering impedances). The resulting "loop flows" on lines far away, even outside the designated control area of a system operator, cause major problems in managing transmission grids. When an energy-balanced injection and withdrawal is charged for losses (energy dissipated as heat) and transmission, these charges represent the total over all paths. The absence of point-to-point transmission has had the economic consequence that property rights are not assigned by title (in contrast, title to gas is tracked continuously, even though it is perfectly homogenous). No one owns power per se; rather, qualified market participants obtain privileges to inject or withdraw power from the network at specific locations.

These privileges bring obligations to comply with technical rules and procedures for settling accounts based on metered injections and withdrawals. Thus, all rights are reciprocal and derive from contracts, typically tariffs or explicit contracts that govern participation in the system. Some parties own generating plants and transmission lines but it is not these properties that are traded in multilateral power markets. Various financial rights are created by contracts, such as transmission "rights" that reimburse usage fees for transmission, and in some cases, allow scheduling priority for the right-holder. Jurisdictions such as the U.S. require open access to the transmission system on nondiscriminatory terms, preclude transmission owners from withholding capacity, and in some areas assign control to a system operator.

Incompleteness of the market would be a minor deficiency were it not that most demand values far exceed supply costs and have large stochastic and cyclic components. Given present 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. With few exceptions, customers' rights to withdraw power from the grid are unrestricted, and some jurisdictions give suppliers comparable rights to inject power that is paid the spot price at the injection point.

This requires strenuous efforts to muster sufficient generation sources and transmission capacity to supply predicted and then actual demand, supplemented by reserves to meet contingencies (reserves are called "ancillary services" in the power industry). 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 subject to Kirchhoff's Laws, the transmission grid is highly complex and vulnerable to instability, cascading failures, or collapse at great cost. Failure of a line or generator can precipitate a crisis that develops orders of magnitude faster than generators can ramp up or down to compensate; other problems such as voltage deficiencies evolve slower but require continuous monitoring and supplemental resources for corrective actions. In general, the maximum cushion available to operators is the ten minutes in which governors and automatic controls on generators can compensate for energy imbalances in the system.

The chief economic consequence of the pervasive externalities and continuous requirements for balancing the transmission system is that within a short time frame it is not feasible presently to rely on spot markets mediated solely by clearing prices. Partly this is because competition is imperfect when the operator needs specific kinds of resources immediately and in particular locations, but fundamentally it reflects the necessity of more and quicker coordination than markets provide. The situation is like other prices-versus-quantities contexts where technical rigidities create complementarities and localized market power that outweigh the advantages of substitution among competing offers in the market. The better alternative relies on direct quantity specifications, with auxiliary rules used to settle accounts ex post. In the case of power, real-time control is managed by a system operator using procedures influenced more by engineering than economic considerations, and invoking directives when markets fail—called "bid insufficiency" or "out of market" in the power industry.

There can be only one spot market for energy, the real-time "balancing market" conducted continuously by the system operator as an integral part of its management of transmission. The spot market is just the first in a cascade of options to balance energy flows and maintain reliability. Offers to adjust energy generation or load are invoked first (thereby altering the spot price), then a hierarchy of reserves ordered by response rates, and finally directives. Exclusive responsibility for technical control of system stability accounts for the operator's unique role in managing the spot market, and precludes competition from other market-makers.

If the spot market were complete and competitive then all forward markets could be organized around financial contracts pegged to spot prices. In fact, however, the sequence of forward markets never attains this ideal. It is difficult to include fully such intertemporal effects as generators' startup costs and ramping constraints and hydro reservoirs' limits on total energy, and such spatial effects as transmission constraints, of which some are local (thermal limits, voltage and reactive power requirements) and others with huge external effects are global (stability, security against cascading failures). The end result in many systems is that the scope of the operator's authority extends over a longer period before real-time to cope with the many implicit coordination tasks and unpriced scarce resources affecting performance. For instance, in some systems the system operator is explicitly charged with authority to ensure physical feasibility of proposed schedules a day or some hours before real-time operations.

An important design issue is thus the scope of the system operator's authority to manage forward markets, which is the topic addressed first. This includes implicitly the regulation, governance, and incentives affecting the operator, but I focus here on designs of forward markets. Hereafter, I abbreviate *system operator* as SO.

2. EXTENT OF RELIANCE ON MARKETS

There is presently no standard organization of liberalized wholesale markets for power. Every jurisdiction uses a different structure of regulation, governance, system management, and markets. One way to place these various structures along a single dimension focuses on degrees of reliance on market processes, as compared to managerial discretion. This dimension reflects differences in the extent of unbundling of energy, transmission, and reserve capacity into separately priced commodities, and differences in the priority assigned to enhancing coordination. Along this dimension, the organizational forms span a spectrum between two extremes.

One extreme assigns broad authority to the SO, intending to recapture advantages of tight coordination obtained previously from vertical integration and cooperative power pools. Examples are in Britain 1989–2001 and in the U.S. Northeast, including New England, New York, and to a lesser extent, Pennsylvania-New Jersey-Maryland (PJM). The other extreme form is more decentralized because the SO's responsibility, beyond managing transmission, is to facilitate private markets for other ingredients. For instance, the SO allocates transmission access, but also it auctions tradable transmission rights and conducts auction markets for counterflows used to ease transmission congestion. Similarly, the SO ensures joint feasibility of operating schedules, but each participant retains discretion to construct his own schedule to meet obligations contracted in various markets. And again, the SO conducts procurement auctions to obtain reserve capacity, but each participant retains options to self-provide or contract elsewhere for resources meeting his reserve obligation. Separating the markets for energy, for transmission rights and counterflows, and for reserves sacrifices tight coordination. But it reflects the priority accorded to maximizing the role of private markets and minimizing decisions immune to market tests. Examples are in Australia, Scandinavia, California 1998–2000, and Texas, as well as Britain's new system that began operation in 2001.

Adequate labels for these organizational forms escape me, but here I call them *integrated* and *unbundled*.[3] Their difference stems from distinct solutions to the tradeoff between tighter coordination and greater reliance on markets. This tradeoff is not intrinsic, since one can envision highly evolved markets with elaborate pricing sufficient to achieve perfect coordination. But in practice the markets in unbundled systems are presently so crude that this tradeoff is a major consideration, and an integrated system could be superior.

[3] Earlier reviewers objected to *centralized* and *decentralized* and another referee proposed SO *commitment* and *self-commitment*.

As mentioned, integrated systems imitate vertically integrated operations. Typically they inherit procedures from national monopolies or regional power pools that previously coordinated the schedules of utilities or distribution companies serving adjacent areas. Their characteristic feature is a "smart market" in which sophisticated optimization software is used to minimize a measure of the cost of serving demand (or maximize gains from trade when demand-side bids are included), subject to both system constraints, such as transmission capacity, and each participant's operational constraints, such as a generator's ramp rate. The aim is to strengthen physical feasibility and ensure coordination of all aspects of energy, transmission and reserves. An integrated market allocates resources according to a coherent plan, and optimization expands the scope and completeness of the market by recognizing operating constraints that represent scarce resources not traded or priced explicitly. Usually the smart market excludes flows from energy trades contracted bilaterally (about 60% in PJM), which are charged the prices for transmission and reserves that emerge from optimizing included flows. In principle, these prices are derived from the optimization's shadow prices on transmission constraints. In general, prices are established only at external boundaries of an integrated system. For instance, so-called nodal pricing sets a price for energy injection at each location; thus, the price for energy transfer between two points—the difference between the prices at these points—summarizes the shadow prices on all resources affected by the transfer.

The simplest unbundled systems rely on a sequence of separate (and sometimes, multiple competing) forward markets for energy, transmission, and reserves, each priced separately. Each price is simply the one clearing that market, and each participant schedules its resources to fulfill its sales or purchases. The reader may want first to read Section 3 where I provide more detailed descriptions of the structure of these markets and their procedural rules.

I discuss these two extremes as though they are dichotomous when hybrid versions might obtain the best of both. The main problems are (a) sustaining incentives within smart markets in which optimization is used to allocate multiple scarce resources and to account for other constraints that are not priced explicitly, (b) enabling market participants to contest the prices derived from this optimization by offering better terms, and (c) taking advantage of participants' superior information about local factors affecting scheduling and operations of their own plants. Designs of both types are converging as innovative solutions to these problems are devised incrementally; and in the U.S. the federal regulator has proposed development of a "standard" design. For instance, several integrated systems allow suppliers to self-schedule their generators, and some unbundled systems use auxiliary optimizations. The exposition examines the two extremes and then sketches a hybrid.

2.1. *Integrated Systems*

Two characteristics of integrated systems are a long-term relational contract among participants, and a smart market that includes overall optimization of

operational decisions. Besides specifying market rules and sanctions, the contract specifies reciprocal obligations. In New England, for instance, these include demanders' obligations to obtain options on sufficient installed capacity, and suppliers' mandatory participation, including obligations to offer all operable capacity, all of which must be available in real-time operations even if not assigned to reserve status. A supplier's participation is voluntary in systems such as PJM, but those who opt to trade in private markets are price-takers, paying the transmission prices determined by those who do participate (other systems, such as New York, allow contingent adjustments in case the price is too high). In general, integrated designs preclude market tests of the SO's decisions and prices, since there are no alternatives. In a typical design, the key economic aspects are:

(a) Forward (day-ahead) optimization of all generation (net of bilateral trades), transmission, and reserves. The optimization includes intertemporal factors such as startup commitments and constraints on generators' ramping rates and reservoirs' potential energy.[4] The resulting schedules are indicative plans, since they are re-optimized on a shorter time frame (hour-ahead) and again in real-time operations.

(b) Pricing and settlements are based on system-wide opportunity costs as measured by shadow prices on system constraints, such as the necessary equality of energy supply and demand in real time, and limits on transmission capacity.

The optimization and settlements use submitted bids to represent costs and values. Some designs use three-part bid formats in which each supplier specifies its fixed cost of startup and minimum running cost, in addition to its schedule of marginal costs. All three cost components are taken into account in optimizing generation sources (with unrecovered costs charged to all participants) in the day-ahead optimization called unit commitment. Other versions using one-part bids require each supplier to plan its unit commitment and to absorb the costs.

The basic argument for thorough integration is that comprehensive optimization is necessary to minimize the total cost of ensuring reliability and coordinating generation, transmission, and reserves to meet predicted demand. That is, productive efficiency requires optimization, an argument that reflects the focus on mustering supply-side resources to match demand. Although the SO operates on a shorter time scale, its role is otherwise much like the "single buyer" paradigm used in countries whose government-owned systems purchase from private firms via long-term contracts. In economic terms the advantage is better pricing of supplies, in the sense that shadow prices derived from constrained optimization more accurately reflect the system-wide opportunity costs of scarce supply-side resources, both intertemporally and spatially.

In terms of organization, optimized operations rely on an SO with exclusive authority to manage the system and to conduct a unified market, including both

[4] 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.

forward planning and real-time operations. The aim is a first-best solution to the problem of minimizing the total cost of serving demand. In its purest form in Britain before 2001 (and the U.S. Northeast still), the market was run as a direct revelation game: participants revealed their supply costs and demand values, as well as various technical constraints, that became inputs to an algorithm. The objective is usually stated as maximizing the gains from trade as measured by these submitted costs and values, or when the demand side is not included or demand is inelastic, minimizing the total cost of serving the predicted load and real-time adjustments to serve the actual load.

Incentives are addressed via settlement rules that specify financial payments. For instance, with nodal pricing a generator is simply paid the spot price at its location; implicitly this is a bundled price for energy net of charges for transmission from its location to a reference location. Full incentive compatibility is never attempted, relying instead on competition or regulation to ensure that settlement prices derived from shadow prices on the main system constraints (supply = demand, transmission $\leq$ capacity, reserves $\geq$ X% of load, etc.) suffice. When competition is weak, integrated systems rely on strictures and sanctions to control abuses, and in the long run, contestability from entrants.[5] Long-term relational contracting enables some internal discipline, but the Market Surveillance Committee in New Zealand may be the only one with sufficient quasi-judicial authority, since such committees in other jurisdictions are advisory. The U.S. regulator is empowered by law to ensure "just and reasonable prices" by requiring bids to reflect actual costs, and regulatory agencies in other countries retain comparable authority, but invoking such powers obviates some of the intended advantages of deregulating wholesale markets.

The counter-argument is that, absent regulatory enforcement, cost minimization is a fiction without stronger incentives to ensure that bids reflect actual costs. In systems like Britain before 2001, where incumbents enjoyed substantial market power, it was sometimes obvious that bids were intended to manipulate the algorithm (OFFER (1999)).

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. The theory of mechanism design offers clear specifications of settlement rules designed to ensure incentive compatibility, the simplest being the Vickrey rule that makes truthful revelation of privately known costs a weakly dominant strategy, absent collusion.[6]

[5] In Britain these methods collapsed in the summer of 1999 as the transition to a new market structure neared and a moratorium on new gas-fired plants was imposed. Prices in the first two weeks of July averaged 80% above the year before, allegedly due partly to suppliers' submission of false operating constraints (OFGEM (1999)). Manipulation of operating constraints has also been a problem in the U.S. Northeast, especially New England.

[6] Additional assumptions are required when privately known components of costs are not statistically independent.

But implementation encounters difficulties that are ultimately political; e.g., a Vickrey auction entails price discrimination favoring suppliers with market power, paying each what it could obtain by withholding supply or inflating bids. Requirements for nondiscriminatory pricing and direct mitigation of market power are often included in enabling legislation, thereby precluding alternatives that promote productive efficiency via incentive schemes.

In contrast, arguments for unbundled 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. Smart markets could consolidate and optimize forward markets, but proponents of unbundled designs doubt the SO needs to conduct these. They argue that the SO's authority to manage transmission and real-time balancing, plus minimal intrusions into forward markets for transmission and reserves, should not extend to forward energy markets beyond assuring physical feasibility. The operator's narrow scope is seen as sufficient for reliable operations, and any greater scope would remove more decisions from market tests. The incentive effects of unbundled markets are diffuse but I attempt to identify some in Section 2.2.

These arguments pro and con support integrated designs in vigorously competitive situations where the gains from tight coordination exceed the gains from stronger incentives. Early implementations retained integrated operations amidst optimism that liberalized markets would be sufficiently competitive to suppress strategic behavior. The initial experience in Alberta, Australia, Britain, and others justified this optimism because incumbents' long-term hedging or vesting contracts induced strong incentives for maximizing output, and thereby low spot prices. Optimism was justified in Argentina because generous capacity payments attracted surplus capacity. It proved unjustified after the contracts expired in Britain, where the market power of dominant firms provoked protracted struggles with the regulator. Most U.S. systems encountered market power problems right from the start; e.g., the New England operator suspended its principal markets for reserves because they were "not workably competitive" (New England Independent System Operator (1999)).

Although productive efficiency is their justification, integrated systems tolerate inefficiencies that must be recognized to obtain an accurate comparison with the unbundled systems described in Section 3 below. Some are prosaic, such as optimization based on imperfect models of generators' operating characteristics, or 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; in fact, to suppress gaming, flexibility for suppliers to revise technical data is limited or excluded. Settlement procedures ignore effects on incentives and gaming. Heavy reliance is placed on directives, sanctions, and penalties when usually the optimal penalty for deviations is to charge or pay the spot price. Integrated systems usually spread unrecovered start-up costs over all participants in the form of an "uplift" charge. I elaborate three more examples.

In some cases prices are related vaguely to optimized shadow prices on scarce resources. For instance, the remuneration paid to reserve capacity in New England is calculated ex post to justify what actually occurred. In PJM's real-time market the operating engineers decide continuously on measures required to maintain transmission reliability; then the price for energy injection at the location of each supplier is set equal to that supplier's bid for the quantity chosen by the engineers. That is, real-time pricing is as-bid for the quantity desired from each generator. Such practices prevent arbitrage by others, who might offer competing bids to alleviate the hidden constraints recognized by the engineers, and thus they preclude market tests of the administered prices. The difference between the injection prices at two locations can be interpreted as the implied scarcity value of transmission between these locations, but it is only by solving a large set of equations that one might infer the implicit shadow prices on the transmission constraints enforced by the engineers. In contrast, unbundled systems are more explicit, and more important, every price can be contested by competing offers. California priced energy and transmission separately, and its revised design in 2000 specified explicitly the local constraints on generation and transmission enforced by the operating engineers, thus enabling accounts to be settled at the clearing prices for suppliers' adjustment bids accepted to satisfy those constraints. The difference between the two designs lies in an integrated system's prerogative to fix prices only at points on the boundary where it interacts with suppliers, internalizing all else, whereas an unbundled system's multiple markets for each of the various resources require explicit specifications and market clearing prices—in the case of California cited above, these are the clearing prices for adjustments sufficient to stay within the specified constraints. The scarcity values of transmission capacities in market designs like PJM are inaccessible to participants, whereas designs that include separate markets for adjustment to satisfy explicit transmission constraints make these prices transparent and thereby reveal the scarcity values of transmission capacities. This is important because generation can substitute for transmission: incrementing and decrementing generation at the two ends of a congested line creates a counterflow that is a perfect substitute for more capacity.

Pricing is especially vulnerable to incentive effects. An example occurs in integrated 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 wrong because the only economically relevant prices are spot prices; that is, it is only in real time that supply and demand must balance physically. In fact, however, markets are imperfectly competitive, and forward markets serve an economic function that is especially important in power markets. Forward markets decide irreversibly which among the operable plants will be committed to run, and constrain the

range of output levels that are feasible later due to ramping constraints. The successes of unbundled designs encouraged several integrated systems to adopt multiple settlements.

Pricing is distorted whenever optimization is imperfect. A typical example is real-time optimization that relies on a 24-hour rolling horizon. Unlike the price derived from the day-ahead optimization, the spot price in an hour of peak demand takes account of intertemporal constraints on ramping down without accounting for the constraints and imputed costs of previous ramping up to meet the peak, so it is biased compared to the price computed day-ahead. The net effect is to undervalue flexible resources used to meet peak loads, and indeed systems such as New England that used this approach were often short of flexible resources (e.g., some combustion turbines were removed and installations of new units were deterred). 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 able to 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 insignificant.

These examples are indicative of a pattern in which unbundled markets are judged deficient because they are incomplete and loosely coordinated, but the incompleteness of optimization models in integrated systems is not recognized—the optimization models are incredibly elaborate in terms of engineering detail but devoid of some salient economic features. Incomplete markets are explicit in unbundled systems and implicit in integrated systems: these limitations are not intrinsic, of course, since they reflect mainly the state of the art for implementing principles from economics and optimization theory. For example, insufficient rewards to flexible resources stem from a missing market for load-following services. Because load variability is not priced explicitly and generators have limited ramp rates, a theoretical model imputes supply prices both to power (the rate of energy production) and the time rate of change of power, but in practice only power is priced explicitly and averaged over a duration as long as an hour in forward markets. Compared to completely unbundled systems that clear 24 hourly markets independently, an optimization that includes ramping constraints improves pricing by inducing more inter-hour price variability, provided multiple settlements are used so that these more variable prices are actually paid to flexible resources.[7]

[7] There are small incentive effects from using shadow prices on ramping constraints in settlements, because for each plant the net payment over a daily cycle would be nearly zero; the incentive effects

2.2. *Unbundled Systems*

So how do unbundled designs fare? Operators might consider it a miracle that they worked at all in systems like California and Australia with thermal generators affected by startup costs and ramping constraints (compared to Norway with hydro reservoirs that can vary energy generation at a moment's notice) and with must-run nuclear units, but evidently they do.[8] California was remarkable because of its complete reliance on voluntary participation except for plants designated must-run for local reliability.

An economist's first response is more sanguine because, in principle, unbundled markets solve the dual of the primal optimization used by integrated systems. The devil is in the details, however, due to two features of current designs.

(a) There is no explicit coordination of the markets for energy, transmission, and reserves. Because these markets typically operate in sequence and clear independently, one needs 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 and consumption.

(b) Intertemporal costs and constraints are not included explicitly and must be internalized by participants; e.g., each bidder self-schedules his plants (startup, ramping, etc.) to generate energy sold in forward markets. Intertemporal 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.

These features stem from limitations on the bid format and on the complexity of the market clearing process in initial implementations. An example of the effect of (a) is that a supplier must submit its energy bid knowing neither the price of transmission nor the price it could get for reserve status. An example of (b) is that a supplier might be unable to supply its multi-hour sales contracted in the day-ahead market, in which case it must sell more or buy some replacement energy in the hour-ahead and/or real-time markets. Thus, problem (a) stems from sequential markets for the three products, but also problem (b) is eased by repeated markets for energy and transmission. I outline three views about the effects of loosely coordinated, unbundled markets on efficiency and incentives.

The first is the sanguine view that these problems are not serious in terms of efficiency. Because the markets are repeated every day with the same participants and little uncertainty (day-ahead forecasts of hourly loads are typically accurate within 3%) it is plausible that bidders in one forward market can anticipate prices

---

are mostly in the increased variability of energy prices. Ramping constraints can inflate market power, as in Britain in the Summer of 1999 when large increases in bidders' use of "inflexibility markers" were interpreted by authorities as enhancing the two largest firms' ability to increase prices, "effective price setting competition is eroded" (OFGEM (1999)).

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

in subsequent markets. Self-scheduling enables a supplier to allocate generation among its several plants to ensure feasibility, or alternatively, the sequence of long-term, day-ahead, hour-ahead, and real-time energy markets provides ample options to remedy physical infeasibilities with later trades. In this view, decisions taken in forward markets can be implemented with ample flexibility, and in any case are reversible in subsequent markets before real-time. Smart markets are thus seen as unnecessary and it is better to focus on sequential adaptation to evolving private and public information in a sequence of simple markets. Incentives are aligned in the theoretical sense that the market outcome is presumably a Nash equilibrium in the bidding strategies of participants—in contrast to an integrated system whose optimization attempts to mimic a Walrasian equilibrium when in fact bidders are not price takers and the outcome is a Nash equilibrium in distorted reports of costs. I doubt that these two Nash equilibria can be ranked in terms of efficiency, but in Section 3 I examine whether incidental effects of unbundled markets improve efficiency by strengthening competition. These incidental effects include additional features such as explicit auction markets for reserves, and most importantly demand-side bidding, including interruptible loads offered as reserves. The reluctance of integrated systems to include some of these features may be a historical residue. Other incidental effects are ones of omission: absent a long-term relational contract binding on market participants, voluntary markets omit installed and operable capacity requirements, offer few devices enabling incumbents to deter entry, little power to impose strictures and penalties differing significantly from spot prices, and usually no capacity payments that subsidize inefficient plants.[9]

The second view recognizes the severe incompleteness of the simple markets used in unbundled systems, but argues that this potential deficiency is largely eliminated by the rich sequence of markets. This view is consistent with theory that reaches a similar conclusion for continuous trading of simple contracts (Kreps (1987)), but its proponents argue in practical terms. For example, knowing the transmission charge when submitting an offer to supply energy is not crucial because in the later transmission market a supplier can submit additional bids to adjust his generation up or down (accepted bids alleviate transmission congestion by creating counterflows, and the SO's transmission charge is the marginal cost of these counterflows). Similarly, commitments in the day-ahead energy and transmission markets do not prevent bidding into the auctions for reserves, since conflicting commitments can be remedied by purchases or sales in the day-of, hour-ahead, and real-time markets. These options recur in the day-of markets, and even in real-time one can renege on previous commitments by paying the real-time price for deviations. This flexibility stems ultimately from voluntary participation, which enables a market participant to adjust commitments repeatedly as the actual delivery time approaches, and indirectly from the favorable incentive effects of multiple settlements, which ensure that each transaction, adjustment, or deviation is charged or paid only the price in the market where it is

[9] Wolak and Patrick (1998) indicates that in Britain the probability of lost load used to justify capacity payments averaged ten times the actual frequency.

contracted. A key consequence is that transactions in forward markets are inherently financial, because the implications for physical scheduling can be adjusted in later markets.

This aspect contrasts with integrated systems that interpret optimized plans as physical commitments and penalize deviations, even though physical feasibility is not binding until real-time. Some shift toward financial interpretations of forward markets has begun as integrated systems increasingly allow suppliers to "re-declare" their costs and to make some adjustments in their schedules without penalties, though these options often intensify gaming. Substantial differences remain because an integrated system's insistence on physical feasibility in forward markets is at odds with the flexibility and essentially financial character of the sequence of separate forward markets used by an unbundled system. A basic issue that divides them is whether early assurance of physical feasibility and tightly coordinated scheduling are more important than potential gains from flexibility and dispersed decision-making enabled by unbundled markets.

This issue would be merely theoretical in a comparison between integrated systems and most unbundled systems, since their performance in terms of physical feasibility has been comparable. During the 2000–2001 crisis, however, California's unbundled system often teetered on the brink of infeasibility, due initially to supply shortages, but severely exacerbated by the utilities' choosing to make insufficient purchases in forward markets so that they could take advantage of price caps in the SO's real-time market. This episode could be an argument for extending the SO's control of forward markets, or an argument against regulatory interventions that distort unbundled markets. Given this and other distortions (such as the "dec game" described in Section 3.3), nevertheless, the SO imposed increasingly stronger controls on forward markets to enhance physical feasibility, and eventually the regulator required that day-ahead schedules resulting from the energy, transmission, and reserves markets must be physically feasible, and large penalties were imposed for real-time trades exceeding 5%. Requirements for physical feasibility in advance of real-time operations moved the California system closer to an integrated system. The trend elsewhere was the opposite, as in new systems implemented in 2001 in Texas and especially Britain, where the new design deferred requirements for physical feasibility to shortly before real-time.

The third view is that the simple unbundled markets in operation now are a transitional step until designs of consolidated forward markets are developed. The central issue is how to conduct such a market without intervention by the SO.[10] The problem lies in the definition and assignment of property rights. When the SO retains full control of transmission capacity, necessarily it conducts the

[10] One can ask why some jurisdictions minimize the operator's role, since consolidated markets conducted by the system operator abound and operate with some success. Some conjecture fear of the long-term consequences if the SO's monopoly power were comprehensive, but the chief factor is that these jurisdictions are mainly those that have no history of coordination via regional power pools. Participants hesitate to agree on terms of a relational contract that is a pre-condition for the principal-agent relationship with the system operator.

only markets for transmission. Then the sole option is for private parties to issue financial instruments that hedge against transmission prices, but markets for such instruments have not developed anywhere near the depth required to enable consolidated markets, apparently because they are too risky presently. In the U.S. this situation changed when the regulator required the SO to issue instruments sufficient to assure transmission "price certainty" for those who buy them. PJM's integrated system issues annual financial instruments that hedge only against point-to-point transmission charges, which are too specialized to sustain active secondary markets, and in fact the only secondary market is the monthly market for re-configuration conducted by PJM itself. But in California's unbundled system the SO auctioned annual "firm transmission rights" that included rebates of day-ahead transmission charges on the major lines, and importantly, scheduling priority over competing requests for access. The scheduling priority makes these instruments functionally equivalent to annual leases of transmission capacity. In turn, this enables existing private markets for energy to be expanded into consolidated markets for energy and firm transmission *rights*, and in that case there is no barrier to optimization in these markets to take account of operational constraints too (see Section 2.4). This third view thus argues that consolidation of forward markets does not require the SO to have an exclusive franchise: if tradable instruments like transmission rights are issued, then private parties can conduct combined forward markets for energy and transmission rights. The source of the SO's seemingly necessary role in integrated systems thus lies in the extension of its exclusive real-time control of transmission reliability to a monopoly on forward trading of transmission rights—which is not necessary even if it is judged desirable to enhance coordination.

2.3. *A Summary Comparison*

One way to obtain an overall perspective on the contrast between integrated and unbundled systems 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 when vigorous competition makes the first-best incentive compatible. Departures from this equivalence differ for the two designs.

Integrated designs start from the premise that, as in traditional power pools, participants are bound together by a relational contract and, in effect, they employ the SO as the exclusive manager of all multilateral markets—forward and spot, energy and transmission. The motive is to realize gains from tight coordination in daily operations, and potentially from longer-term obligations and subsidies aimed initially at strengthening overall reliability. Problems arise because the incentives for participants to cooperate are undermined. Manipulations by participants with market power cause problems because few instruments are effective counter-measures; in particular, pricing and settlement rules sufficient for incentive compatibility are too complex to be practical and often entail price discrimination, 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 distorted when detailed knowledge of participants' costs and values is replaced by submitted bids in limited formats, and impaired further when models and algorithms restrict flexibility and distort prices. The SO's decisions, and ultimately, prices are immune to market tests.

These considerations imply that integrated designs are most effective when there is vigorous competition, or if competition is limited, when there is either regulation or a legal cartel with ample powers of enforcement as in New Zealand. Their advantages are greater too when optimization to meet system constraints is more important than participants' flexibility to optimize their own operations, and shadow prices on system constraints are more accurate measures of opportunity costs than clearing prices in markets.

Unbundled designs start from the opposite premise that participation is voluntary, with no long-term relational obligations other than a general tariff approved by the regulator, and that competing forward markets are encouraged to the extent feasible. The necessity of a system operator with exclusive authority to manage the public good represented by the transmission system is acknowledged. The SO's responsibilities include real-time operations that protect system reliability, but its authority to intervene in forward markets is limited to cases where prior commitments promote reliability, such as day-ahead scheduling of transmission. The motives for limiting the scope of the operator's authority are to isolate its monopoly control of transmission from competitive energy markets, and to enable unbundled pricing of energy and transmission. This constraint on the scope of the SO's role can be binding and impair efficiency when forward markets are severely incomplete, poorly coordinated, or distorted by regulations. Such a conclusion might apply to the sequence of simple markets implemented in current designs, but the matter is inconclusive because ample flexibility and repeated trading opportunities might suffice to simulate complete markets and improve coordination.

Obviously, the case for integration is strongest when there is vigorous competition to sustain incentives and optimization is accurate enough to imitate complete markets. Equally, the case for unbundling is strongest when markets are competitive and when trading opportunities are rich enough to approximate complete markets. This similar comparison of polar opposites indicates practical criteria: how closely does a proposed design approximate complete and competitive markets. Although power markets are never complete nor perfectly competitive in the Walrasian sense, design efforts in a pragmatic vein can focus on these two dimensions. In particular, the differing modes of competition and price determination in integrated and unbundled systems suggest that the choice hinges on whether relational contracting and unified management by a system operator induces as much competitive pressure as voluntary day-to-day trading in forward markets offered by competing parties. Similarly, the deficiencies of those optimizations used in practice can be compared with the residual incompleteness of unbundled forward markets.

The choice between these approaches has not yet had a crucial test because implementations have been limited to small jurisdictions. The regulator in the U.S. required that by 2002 the many local power systems in the U.S. be consolidated into as few as four regional transmission organizations (e.g., the entire West). Perhaps operations in Europe will eventually be organized on a similar scale, as Scandinavia and Australia are now. As the scale increases, it becomes harder for an integrated system to optimize everything simultaneously. The key task is to coordinate operations on the two sides of each seam between smaller control areas. It is unclear now whether this is better done by an integrated system that optimizes operations on a regional scale, or by using markets to establish prices for transmission across boundaries. It seems likely that markets for trades between areas enhance prospects for greater use of markets within each area separately.

2.4. *Hybrid Systems*

Resolution of these tensions lies in hybrid designs that enable coordinated markets. To illustrate, I sketch how forward markets for energy and transmission rights can be unified to capture gains from tighter coordination (Chao et al. (2000)). The key feature is a smart market in which prices and resource allocations are obtained from a constrained optimization. This can be done even if the SO's scope is only to manage transmission and the balancing market, provided firm transmission rights (FTRs) are issued in sufficient amounts. Although SOs like to hoard some transmission capacity until the last moment, simplify by supposing that FTRs for transmission across each major interface are auctioned annually for all the transfer capacity potentially available in each direction. Each FTR has a par value (1 mega-Watt) that the SO adjusts daily to account for current conditions. The SO also publishes the matrix of transfer factors used by its engineers that day: each factor specifies the fraction of the power injected at any point and withdrawn at a reference point that the SO predicts will flow through each interface.

With these ingredients available, privately organized forward markets have many options for coordinating allocations of energy and transmission rights. Energy and transmission rights can be offered as bundled products in bilateral markets. A power exchange can conduct a smart market for energy and transmission rights that also includes generators' operating constraints, such as ramp rates and minimum production rates, and auxiliary costs for startup and running—indeed, all aspects of unit commitment and scheduling that integrated systems keep within the SO's control. Of course the SO retains authority to adjust schedules to ensure zero net flow on each interface other than the amount allocated to FTRs. The FTRs provide financial hedges against transmission charges, and they include scheduling priority that provides some protection from curtailments.

Actually, a fully consolidated market for energy and transmission rights is not necessary. The market for FTRs can occur after the close of the energy market. Suppliers in this market are owners of FTRs and they submit supply functions.

Demanders are participants in the previous energy market and each submits a demand function indicating for each price the quantity it wants to hedge against the usage charges determined in the SO's later market for adjustment bids to alleviate congestion. The FTR market operates by optimally matching a demander who wants to hedge an injection at one location with one who wants to hedge an extraction at another location. The *sum* of their bids is their joint bid for an FTR that assures transmission between them. As usual the market is cleared at the prices that equate supplies and demands of FTRs for each interface. The price paid by each demander is the shadow price on a marginal extraction at its location, or its negative for an injection, so that the difference in prices at two locations is exactly the price of an FTR purchased to assure transmission. The two demanders might be the same, which occurs when that demander provides a counterflow that substitutes for an FTR, and thus competes directly with the SO's market for adjustment bids.

The essential point is that the priorities that motivate integrated and unbundled designs do not necessarily conflict. The seeming conflict between tight coordination by the SO, and contestable markets managed by other parties, is an artifact of insufficient contracts. Others can conduct smart markets for forward contracts in energy and transmission jointly if the SO auctions sufficient FTRs that can be traded in secondary markets. Similarly, others can conduct markets for resources to fulfill reserve obligations if the SO allows participants to self-provide or purchase them. It may be that vertically integrating the SO obtains some further advantages that additional contracting cannot, but if so then these must be compared to the stronger incentives and market tests of unbundled designs. My guess is that unbundled designs will ultimately include contracts and markets sufficiently rich to enable all the spatial and intertemporal factors relevant for forward planning to be included. If the SO can specify these factors explicitly via operational constraints in an optimization, then eventually there will be contracts and markets for resources that relax those constraints.

Admittedly this kind of hybrid differs greatly from the "standard" design proposed in the U.S., which is based on PJM design in which the market is bifurcated: those who volunteer for optimized scheduling by the SO account for 40% of the market, while the remainder (including those utilities who remain vertically integrated) rely on bilateral contracts, self-schedule, and pay prices for transmission and reserves derived from the optimized segment. And in Britain the new system relies almost entirely on bilateral contracting.

2.5. *Comparison with the Gas Industry*

The extent of integration is contentious in other industries previously dominated by vertically integrated monopolies. In Victoria and Britain, a system operator manages transmission of natural gas, whereas in the U.S. each interstate pipeline owner manages its own system, subject to regulations requiring monthly resale markets for firm transmission and daily auction markets for interruptible transmission. Although the time scale differs, electricity and gas are similar

in that both are homogenous commodities and transmission is based on displacement (electrons or atoms injected are not the same as the ones extracted elsewhere), so a system operator can clear markets by setting prices at injection/extraction points. The price difference, which is the charge for a point-to-point transmission, can be derived as the sum of the shadow prices on the scarce resources used along the route. The U.S. system, however, permits pipeline owners to exercise market power by discriminating at any price below a regulated maximum (resulting in rates of return as much as double the allowed rate on which the maximum price is based in infrequent rate hearings). Each price is for a point-to-point balanced injection and extraction that need not bear any relation to the scarcity value of the resources used, nor does an owner need to identify its scarce resources (compressors, choke-point capacities, etc.).

The advent of new designs elsewhere enables comparisons with the U.S. system. Also, the consequences of allowing transmission owners to conduct their own markets, instead of assigning the task to a system operator, can be studied by comparing the existing designs for gas and power markets in the U.S. Perhaps the closest parallel to the U.S. gas transmission system occurs in the new system in Britain. The SO is the sole transmission owner and operates under an incentive scheme of performance-based regulation administered by an office responsible for both the electricity and gas markets. The SO has substantial discretion to enter into bilateral contracts for balancing energy, to manage transmission congestion, and to procure reserves. Its operating rules in the "grid code" enable it to intervene against behavior considered inconsistent with the spirit of the code.

3. MARKET MICROSTRUCTURE

This section examines in detail the sequence of forward and spot markets in an electricity system, and the connections among unbundled markets for energy, transmission, and reserves. It begins with the spot market where all aspects are consolidated, and then works backward through the various forward markets.

3.1. *The Spot Market*

First is a brief technical summary of real-time operations. As mentioned, power is a flow so operations are continuous. The spot market approximates continuous operation by revising prices every few 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, there are 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 (I ignore other factors, such as provision of reactive energy). Momentary imbalances are detected and corrected automatically by the first reserve category, called regulation, which is provided by dispersed generators equipped with governors and speed controls that respond to frequency sensors. Regulation provides a cushion for about 10 minutes, after which generators must return to prior operating levels to provide regulation services later. As regulation capacity nears exhaustion, its role is replaced by offers in the spot market. The spot market suffices in some integrated systems with central control of all dispatch, like PJM, and in some unbundled systems with very liquid spot markets, like Australia, which rarely purchases options on additional reserve capacity. However, most systems acquire options on reserve capacity in advance (some annually or monthly, others in day-ahead auctions) in amounts specified by established reliability criteria.

When offers in the spot markets are insufficient, and especially when the operator needs to increment or decrement generation in specific locations, the next step is to exercise options on reserves in several categories with successively longer response times. Operating reserves include spinning and nonspinning reserves with response times of 10 to 30 minutes. As options on operating reserves are invoked, options on replacement reserves with response times of 30 to 60 minutes are called to sustain the required margin ($\approx$7%) of operating reserves.[11] Within each category the options are used in merit order according to marginal cost as bid when the purpose is to alleviate system-wide energy imbalances. An option is invoked out of merit order (or if necessary, an out-of-market directive is issued) when needed to remedy violations of reliability constraints at particular locations in the transmission system; in this case, the merit order is adjusted to account for effectiveness in addressing the problem. A further category called reliability-must-run is usually contracted long-term and scheduled in advance to ensure voltage support, stability, or security at key locations.

Each reserve 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 forward markets. Thus, in a stable situation the unused supply incs in merit order represent the *extra*marginal segment of the short-run supply curve at prices above the current spot price, and the unused supply decs represent the *infra*marginal segment at prices below the current spot price. Incs and decs from demanders have the opposite interpretations; e.g., a demand dec is functionally equivalent to a supply inc.

In economic terms, the end result is a continually adjusted real-time price for energy. The SO absorbs the costs of options exercised out of merit order or

[11] A reserve unit is nonspinning if it is not synchronized with the transmission grid. Hydro units and combustion turbines provide quick response nonspinning reserves. The regional reliability councils have differing reserve requirements, and hydro resources are allowed smaller reserve margins.

out of market to maintain reliability, so for settling accounts when there is no congestion, the system-wide real-time price is defined as the highest price among those offers accepted in merit order to balance energy. Distinct prices apply in regions isolated by transmission constraints, and for some purposes prices are further divided into inc and dec prices. It might seem that such a market exemplifies the ideal studied by theorists, but practical aspects intrude.

In a fully integrated 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 dispatchable 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.

Fully unbundled systems operate differently. From the SO's viewpoint, reliability seems precarious because participation in forward markets for reserves and the spot market is voluntary, so insufficient offers of reserve capacity and of incs and decs in the balancing market could jeopardize real-time operations. Other effects of incomplete forward markets are subtler. Forward trades on an hourly basis do not fix output rates within the hour, so rapidly changing demand within an hour, such as the initial morning ramp up, is 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 integrated systems do, but another part is the greater volatility of unbundled markets. In a fully unbundled system the SO does not control or direct dispatch except via the inc/dec and reserve 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. Because arbitrage along the sequence of forward and spot markets is necessary to keep their prices linked, ideally as a martingale, large deviations are penalized only when they might cause market failures.

From this overview of real-time operations and spot markets it appears that unbundled systems are inferior. The SO in an integrated regime can re-optimize the entire system every few minutes to re-dispatch all feasible resources, whereas in an unbundled 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 unbundled 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. Integrated systems require participants to maintain sufficient installed capacity and to offer all operable capacity day-ahead, thus enabling the SO to allocate any portion to reserve status or to dispatch in real time. In unbundled systems the SO conducts

a daily auction to procure sufficient reserves, but the end result could be the same.

What then are the purported advantages of unbundled markets? 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' day-ahead bids to assign some to reserve status, 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, even to the extent of recalling exports.

In contrast, participants in unbundled 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 reserve markets, but progress is slow, due partly to the initial expense of installing adequate 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 the reverse), and similarly spinning reserve can substitute for nonspinning, et cetera. This implies that all reserve markets must be cleared simultaneously, with the result that prices decline as response times increase. In California and elsewhere, initial implementations established a separate auction for each reserve 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 unidirectional substitutability is a lesson in the practical difficulties of implementing markets for multiple goods, even when the theory is clear and simple. California went even further with its "rational buyer" design in which bids were accepted to minimize the total cost of all reserves, even if occasionally that entailed higher prices for slow resources and resulting incentives for suppliers to distort their bids (e.g., offering capacity for spinning reserve that is capable of providing regulation).

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 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 the SO's 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 (Chao and Wilson (2002)). 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.

Even though the complications described above have solutions, reserve markets are a weak link in both integrated and unbundled 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—several 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. Two 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, although there are well developed theories of how to include such losses in transmission prices; cf. Chao and Peck (1996)) and that energy flows along alternative paths obey Kirchhoff's Laws, so they are largely uncontrollable in systems with alternating current. A key feature is that substantial excess capacities of transmission and generation must be held in reserve to avoid cascading failures.[12]

An uncongested transmission system resembles a reservoir to which one can add or subtract water, so in effect it unites all suppliers and demanders in a single marketplace. Many grids are constructed 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 competition and contestability require sufficient transmission capacity. When this is accomplished by building ample capacity, an access fee is charged to recover construction and maintenance costs.

A transmission link is congested when net demand exceeds its safe transfer capacity. Remedies include reducing demand in the congested direction, and creating counterflows in the opposite direction; either reduces the net flow in the congested direction. In unbundled systems, the SO alleviates congestion mainly by scheduling counterflows. These are obtained by selecting among participants' adjustment bids, using incs on one side of the congested interface and decs on the other, as described further below. The access fee is augmented with a usage charge (the price of transmission across the congested interface) that is the SO's marginal cost of counterflow. Integrated systems reduce flows or produce counterflows by directing various generators to contract or expand energy output, providing compensation based on their standing bids for supplying energy.

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

As mentioned, unbundled markets rely on incs and decs to alleviate congestion, which I now explain in more detail. 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

[12] Cascading failures are less a threat in gas transmission. 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. Pressure is maintained by compressors, and flows are directed by valves. Some reserve can be obtained by varying the pressure in the pipe. Long-term reserves are provided by underground storage.

is to raise the price charged in the importing zone for withdrawing power, and to reduce the price paid in the exporting zone for injecting power, until the net flow matches the available 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 *infra*marginal 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 integrated systems, unbundled markets for energy, transmission, and reserves operate in sequence. California's transmission market was similar, but in keeping with its pervasive theme of voluntary participation, it allowed bidders to submit incs and decs to the transmission market that might differ from their bids in the previous energy market. Sometimes the SO received insufficient offers to alleviate congestion and the market failed to clear, in which case a default usage charge was imposed. The default charge was partly punitive, but also it was 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 unbundled designs.

Those systems that impose usage charges only between large zones reflect compromises among competing objectives. Usage charges 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 SO's intrusions into forward markets for energy, due to apprehensions about inherent monopoly power derived from its exclusive control of transmission. This stems from the practical consideration that nodal pricing or an equivalent system of injection charges is presently 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 interfaces are congested; e.g., NordPool uses zones that change daily to conform to the pattern of congestion. 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. Day-ahead zonal pricing also serves as a mutual insurance scheme among participants within each zone, since intrazonal congestion is more sensitive to events close to real-time.

However, this compromise creates adverse incentive effects. Zonal pricing in an unbundled system like California's enables strategies like the following—called the dec game. A supplier who anticipates *intra*zonal congestion affecting his injection node can sell a quantity $3Q$ 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 $2Q$ 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 (even negative) when his dec is invoked out of merit order. The net result is that the supplier collects a profit $[P - p^*]2Q$ or even $[P - p]2Q$ on the extra quantity $2Q$ that he knew initially he would not produce. The adverse consequences could be long-term if anticipated profits from the dec game induce an entrant to build a new plant in the most congested area, the opposite of what is required for efficiency.

The dec game is possible when the transmission market is incomplete. Unlike the injection charges used in nodal pricing, day-ahead zonal pricing charges only for transmission across congested interfaces between zones. The SO's cost of alleviating residual intrazonal congestion in real-time is spread among all participants via its general access charge. In terms of incentives, the basic deficiency is that the SO pays for the incs and decs it accepts to alleviate congestion, rather than charging for causation of congestion. In contrast, the outcome of the SO's day-ahead market for adjustment bids is a charge for transmission across each interface that is the clearing price for counterflows sufficient to alleviate interzonal congestion—the SO need not pay for the incs and decs it accepts because each participant sees that complying with the SO's recommended adjustments is cheaper than paying the charges if he reneges on the incs and decs he offered. The dec game disappears if the SO imposes the analogous procedure in real-time: rather than paying for accepted incs and decs, the SO charges the marginal cost of counterflows to alleviate intrazonal congestion, and each participant pays this charge for the fraction of its flow through the congested line. The net result is largely equivalent to the injection charges of nodal pricing, but the total transmission charge is separated into day-ahead charges for interzonal transfers and real-time charges for intrazonal transfers. This separation reflects the distinction between predictable large-scale congestion between zones, and erratic small-scale congestion within zones. Even if intrazonal congestion is predictable day-ahead, a participant's anticipation of real-time charges for intrazonal congestion deters the dec game.

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 pipelines. In fact, the U.S. regulator 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 might be due to genuine risk aversion, but mainly it reflects marketing advantages obtained by brokers who bundle transmission rights with energy transactions in bilateral contracts.

A financial right entitles a buyer to a continual refund of the usage fee whether or not he transmits energy. When the right includes a scheduling priority, physical access is virtually assured. In integrated 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 unbundled systems like California, each right pertained to the interface between two zones and included both a financial hedge and scheduling priority, which together amounted 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 is less effective due to the rights' absolute priority for scheduling. Recent research also predicts that hedges against transmission fees can magnify the market power of suppliers in import zones (Joskow and Tirole (2000)).

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 Nord-Pool'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, not necessarily progress, is evident in Britain's switch from a central exchange to private 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 less than 4 hours advance notice of energy transactions. Such short horizons are possible because in principle the SO accepts only balanced schedules in which energy injections equal withdrawals,

so it is only in real-time operations that the SO must cope with imbalances.[13] 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.

This sequence of day-ahead, then real-time, operations for the SO meshes with longer time frames in the energy markets.[14] 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. Supplies contracted long-term might pass through the day-ahead market, but they have no effect on market clearing prices because each contract specifies equal amounts supplied and demanded. Contracts are often specified as contracts for differences in which the parties mutually insure each other against the difference between their contracted price and the market price.[15]

Because integrated systems consolidate all energy markets, the basic structure of the forward markets is better described in terms of an unbundled 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 impractical. 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

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

[14] 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.

[15] 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 toward more short-term trading as electronic communication and controls improve to allow more demand-side responsiveness to spot prices.

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 nonstorability of power.

- 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.

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 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. Exchanges are often established as nonprofit entities by legislation or regulation that confines their scope to short-term markets, although a few conduct supplementary markets for longer-term hedges against the exchange price. Their public purpose is to ensure a transparent and liquid forward market whose prices can be used as benchmarks less volatile than spot 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, Britain established one of the first day-ahead exchanges in 1989 and then abolished it, relying entirely on bilateral transactions in private markets. The exchange in California collapsed in 2001 when its trading volume shrank after new provisions allowed utilities to contract bilaterally. Even though other exchanges from Scandinavia to Australia and New Zealand have successful records, the necessity and viability of exchanges remain doubtful. California required its power exchange to compete with bilateral markets and another 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 wither, 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 generalizations 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 hedges 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 U.S. market in the Midwest due to domino effects on other parties, including bankruptcies). A possible 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. 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 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 (Wilson (2001a)).

The deficiencies of existing procedures in exchanges are obvious to economists. The bid format and market clearing procedures take little or no account of intertemporal and spatial factors, and rarely are contingent contracts 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.

Their main advantage is that every price can be contested. Compared to integrated systems with optimized dispatch, market participants have more opportunities to improve a proposed allocation, or to offer better terms than the proposed price. Thus, if the decision is the price paid for energy supplies, then each supplier has an opportunity to offer a lower price, and equally, each demander can offer better terms by bidding to curtail demand. Similarly, if the decision is the price charged for transmission across a congested interface, then each

trader has an opportunity to ease congestion by offering counterflow at a better price. Other 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. Clearing prices are derived transparently from bids with no opaque model and arcane algorithm intervening to compute shadow prices.[16]

4. ALLOCATION OF RISK

State-owned enterprises have the advantage that they share financial risks among all taxpayers. In the era of vertically integrated utilities, they too were effective shock absorbers because their own generation and transmission sufficed for most retail loads. External shocks to hydro supplies or fuel prices were moderated by long-term procurement contracts, and by regulations allowing fuel costs to be paid by retail customers via amortized charges. In addition to buffers inherent in vertically integrated operations, the implicit "regulatory compact" that guided regulation of investor-owned utility companies in the U.S. was an elaborate risk-sharing arrangement. On investments judged prudent ex post by the regulator, a regional utility with a monopoly franchise was assured a rate of return sufficient to obtain funds in capital markets, in exchange for undertaking the obligation of universal service at prices set by retail tariffs. Because regulators approved tariffs periodically, cost shocks and volatile wholesale prices were averaged and spread over long periods, and further moderated by cross-subsidies among large segments of customers. This scheme survived large fuel-cost shocks and high costs for nuclear plants, but ultimately the disparity in some states between the utilities' costs and the prices offered by independent power producers motivated reconsideration.

Federal law after 1978 required states to allow generation by nonutility firms, and technical progress enabled entrepreneurs to compete effectively via smaller and more efficient gas-fired plants. As early as 1983, Joskow and Schmalensee argued in *Markets for Power* that economies of scale in generation had diminished sufficiently to make competitive markets for generation feasible. The regulatory compact was increasingly unstable in the 1980's as utilities lost base-load industrial customers to co-generation and independent generators, and states with high-cost utilities feared loss of commerce and industry to states with lower costs. In the meantime, state regulators saw competitive markets elsewhere (Alberta in Canada, Argentina, Chile, Norway, Victoria in Australia, and especially the

[16] A peculiarity of integrated systems in the U.S. is that detailed models and software are proprietary and confined within the SO—even market participants are unable to replicate exactly how the market "works." The England and Wales system's GOAL program was the opposite: details of its operation became the basis for some contracts.

England and Wales system in Britain) as possible models for "restructured" markets. The motives for establishing power markets varied substantially in these other jurisdictions, and aside from Britain's swift conversion to a wholesale market in 1989, these markets developed slowly and carefully; e.g., Scandinavia's NordPool and Victoria's VicPool designs were revised periodically before expansion into larger regional markets. In the U.S. the impasse broke when the California regulator in 1993 proposed restructuring as a possibility, and in 1994 initiated proceedings to accomplish it, leading to enabling legislation in 1996 that materialized in filings (unanimously approved by stakeholder groups) with the federal regulator in 1997 and initial operations in 1998—in the meantime other states had moved quickly too, so actually PJM began operations earlier in 1998. Omitting the tangled history of these and subsequent events, I concentrate here on two aspects: the allocation of risk implied by the enabling legislation, and the consequences in 2000–2001 when California encountered a severe shock.

4.1. *The California Legislation*

In principle, the legislation established the ingredients envisioned by Joskow and Schmalensee (1983): competitive wholesale markets for energy, an open-access transmission system managed by a system operator, and competitive retail markets, leaving the regulated utilities with responsibilities for distribution and a universal service obligation as the default provider, procuring supplies as needed in wholesale markets. The SO allocated transmission, procured reserves, and conducted the real-time balancing market to protect reliability while a separate power exchange (PX) managed forward markets for energy—although other market-makers could compete with it. Some of the ingredients described in Section 3 were controversial: organizational separation of the SO and the PX and their governance by boards of stakeholder representatives; unbundled prices for energy, transmission, and reserves; reliance on clearing a sequence of simple markets rather than optimized unit commitment and dispatch with locational prices derived from shadow prices on system constraints.

The legislation's allocation of risk stemmed from preoccupation with the three utilities' past and future roles. The key feature was that each utility was allowed four years to apply its net revenues from a sales tax called the "competitive transition charge" to recovery of its "stranded costs" from prior investments and obligations. Until this transition was complete, retail prices were capped at 10% below previous levels; the cap was seen as customer protection in an uncertain future, the 10% as reward for accepting a plan whose main proponents were the utilities and industrial customers. In terms of risk allocation, the retail price cap implied that utilities absorbed the entire risk of volatile wholesale prices. But the incentive effects were good: a utility recovered stranded costs faster and more surely within the allowed transition period if it reduced its procurement costs. To exempt the utilities from prudency reviews of wholesale purchases, they were allowed to purchase only in the transparent markets of the PX and SO where daily prices could be presumed competitive; in addition, this restriction

avoided reviews of bilateral transactions and allegations of self-dealing were a utility to contract directly with its affiliated generation company. Indeed, two utilities' affiliates retained ownership of hydro and nuclear generators, but to enhance competition the regulator proposed divesture of at least 50% of gas-fired generation. The utilities chose 100%, perhaps to speed recovery of stranded costs, and in fact the plants sold for multiples of book values. Rather than using these funds to diversify, or to hedge with investments in facilities in the supply chain, the utilities' parent corporations invested in generators in other jurisdictions—none as capacity constrained as California.

The result was that until the transition period ended the utilities bore all risks of higher wholesale prices. Retail prices were capped and effectively fixed, and the transition charge applied equally to all retail service providers, so little retail competition developed, leaving the utilities obligated to serve nearly all retail demand, at whatever wholesale prices emerged. Wholesale prices were predicted to decline, so the risk of higher wholesale prices seemed small, and would affect only the amount of stranded-cost recovery. There was little concern that mandatory purchases in the PX and SO's day-ahead and spot markets excluded long-term forward contracts as hedges against this risk, and that financial hedges were discouraged by prudency reviews ex post. Proposals to allow hedging were rejected lest the political deal in the legislation would be upset, and because hedging contracts would require prudency reviews.

4.2. *The California Crisis 2000–2001*

With hindsight one sees that the risk of higher wholesale prices was significant. During the 1990's, investments in new generation, transmission, and gas pipeline capacity were essentially zero in California and small throughout the West, and the average age of plants exceeded thirty years. While the economy grew vigorously, the reserve margin shrank steadily from nearly 20% toward the 7% required in hours of peak load, and California depended heavily on imports from hydro generators in the Northwest and thermal plants in the Southwest via transmission lines that were often congested.

The crisis began with drought in the Northwest that curtailed imports into California in the summer of 2000 and pushed gas-fired thermal generation to its limits. Reserve margins dropped to emergency levels in each of four heat waves, especially in Northern California where limited transmission from the south prevented greater reliance on imports from the Southwest. Wholesale prices rose steeply, engendering vigorous criticism of suppliers' apparent market power and attempts by the SO to cap wholesale prices—the Appendix sketches some of these events. The summer crisis deepened in the autumn and winter as the Northwest imported power from California to meet its heating load and thermal generators required downtime for repairs. Wholesale prices rose again as the prices

of emission allowances and California's only fuel, natural gas, reflected scarcities induced by heavy reliance on thermal generation.[17]

California's experiment with restructuring ended in January 2001 when the state intervened to purchase energy supplies because the credit of two utilities was exhausted, and the PX ceased operations after the federal regulator revoked its tariff. What started as a deficit of hydro supplies, then scarcity of thermal capacity and fuel supplies, became a financial crisis when the default providers' credit was depleted.

For present purposes, the important lesson from the California crisis is the crucial role of risk allocation in restructured power markets. In most other industries, risk bearing is spread along the supply chain via long-term forward contracts or financial instruments for hedging. This is optimal since a seller and a buyer have common interests in mutual insurance against the volatility of spot prices. In electricity too it is standard practice everywhere that spot markets account for small fractions of transactions. Observers are unanimous, therefore, that one flaw in California's legislation and regulations was excluding the utilities from long-term contracts. This flaw was crucial because the utilities divested all their gas-fired generation, which could have hedged against hydro shortages, and they relinquished options on gas pipeline capacity. A second flaw was perhaps the cap on retail prices during the transition, since it eliminated customers' financial incentives to curtail demand. However, the smallest of the three utilities emerged from the transition period shortly before the crisis, and when it passed high and volatile wholesale prices to retail customers, a political firestorm erupted—retail customers also wanted level payments as insurance against price volatility.

California's restructuring relied on vague presumptions that competing providers of retail services, if not the utilities, would offer hedging contracts for retail customers, and that ample retail competition would obviate intervention by the state regulator to adjust rates to wholesale costs. But prospects for developing retail competition were killed by imposing the transition charge and the price cap. It was unrealistic to suppose that, when the utilities refused, other service providers would step forward to bear risks of the magnitude of the growing California crisis; indeed, those few with customers abandoned them, taking advantage of the utilities' default service obligation. The state regulatory agency, charged with protecting retail customers, delayed action for many months, until after the largest utility filed for bankruptcy.

The transition charge was intended to facilitate the utilities' recovery of stranded costs. This heavy tax on current sales was also the rationale for the retail price cap (to protect consumers) and the prohibition against long-term

[17] In the U.S. gas pipelines are federally regulated but the commodity market is not. In one view the risk least anticipated was the cascade of events from drought in the Northwest to huge price differentials between the ends of pipelines from the Southwest into California. The regulator investigated allegations that one pipeline favored its energy affiliate's actions to withhold leased capacity, and rationed other shippers to whom firm transmission rights had been sold in excess of actual capacity. Joskow and Kahn's (2001) study of market power includes estimates of the effects of higher prices for gas and emission allowances on generators' marginal costs.

contracting (to prevent self-dealing). The crisis defeated these intentions, jeopardized the financial viability of the utilities, and ultimately required the state to buy power with general funds.

4.3. *Lessons from the California Experience*

If competitive power markets are viable at all they should be resilient enough to survive supply shocks of the kind and magnitude initiating the crisis in California. In the same period comparable shocks affected Brazil and New Zealand among others, and in the American West droughts are frequent. Low reserve margins in California led inevitably to higher spot prices for power and fuel to elicit more generation from thermal units, but the ensuing financial collapse of the utilities was not inevitable. For instance, while the California utilities neared bankruptcy, wholesale spot prices in the Northwest were substantially higher but did not cause financial crises because spot transactions accounted for small percentages of utilities' purchases. The distortions adopted in California to further utilities' stranded-cost recovery while wholesale prices were low included no protection against insolvency when wholesale prices far exceeded the retail price cap. Quite apart from matters of productive efficiency in coping with the initial supply deficiency, the massive transfer of funds from the utilities (and later the state) to power and fuel suppliers induced by higher spot prices was a pecuniary effect that ultimately had huge externalities when the utilities with default-provider obligations neared insolvency. No such externalities occurred when gas markets were deregulated nationally and in California because the utilities were allowed to pass their procurement costs to customers.

A central lesson from the California crisis is that financial arrangements for efficiently allocating risk are important ingredients of restructured power markets. There are other lessons that seem obvious but in fact are difficult to apply. One is that liberalizing wholesale markets should be deferred until there is ample generation and transmission capacity, and account must be taken of the inevitable cessation of new construction while investors wait to see what new regulatory and market structure will emerge after the several years typically required for legislation and implementation. In retrospect, California's restructuring when reserve margins were low and shrinking, even with growing dependence on imports, seems reckless. The difficulty is that often liberalization is undertaken mainly to stimulate private investments in new capacity; thus, several countries encountered the dilemma that the stimulus for liberalization was a growing shortage of capacity that could inflate prices during the early years of competitive markets. Vesting contracts for supplies at fixed prices over several years are often included in the terms of sale of privatized plants to ease this problem.

Another lesson is that the retail sector must be prepared for the downstream consequences of competitive wholesale markets. This is easy if, as with gas, customers tolerate the principle that utilities can pass through their costs of wholesale purchases, presumably leveled or amortized to avoid short-term volatility. Utilities can develop tolerance by offering a variety of retail options that induce

price-responsive demand behavior: real-time pricing, peak-load pricing, service that can be cycled or curtailed when the spot price is high, nonlinear tariffs that encourage efficient usage, incentives for energy-efficient appliances and on-site generation.[18]

In some cases retail competition is initiated directly, as in Texas where each utility must make a portion of its native load available to competition from other providers. Ultimately, restructured wholesale markets require restructured retail markets, since otherwise there is no demand response at the end of the supply chain to price changes upstream. The feature in California that killed retail competition—forcing entrants to offer essentially the same retail prices as the utilities—was intended to enhance the utilities' recovery of stranded costs: surely a salient lesson is that devices for financial transfers to account for past investments should not distort the market design. The transition period in which wholesale and retail markets were encumbered by provisions for stranded-cost recovery was the basic source of the system's vulnerability to financial collapse when wholesale prices rose. A sharp transition to unfettered wholesale and retail markets seems better, even if it means that past obligations cannot be repaid by taxing sales in the new markets.

Beyond these obvious lessons there is the deeper problem of restructuring the regulatory compact. Market liberalization ends the elaborate risk-sharing arrangement in which vertical integration obviates the pecuniary effects of transfer prices along the supply chain, and the regulator assures full recovery of factor inputs and capital investments via tariffs that largely immunize retail customers against price volatility.[19] This arrangement might be replaced by a system of bilateral financial contracts between suppliers and demanders along the chain, from fuel suppliers through to retail customers, and indeed that is typically the intended result when liberalization is preceded by complete privatization or divestiture of power sources. These good intentions cannot be realized, however, unless contracts with the regulator clarify the obligations of default providers. The contract could allow the distribution company to pass procurement costs directly to customers who bear all risk (or better, performance-based regulation enables the company to bear some risk); at the other extreme it could be sold at auction so that it is the franchisee who bears risks; and in between it could be that all service providers share funding of universal service. In California the "contract" with the utilities was mostly implicit so the crisis initiated a continuing struggle over whether retail rates would be raised to curtail demand, and who would bear the accumulating financial deficits (complicated by the largest utility's recourse to bankruptcy court where the judge had substantial powers).

The naive view that implicit understandings for default-provider remuneration might suffice ignores the reality of wholesale power markets. Most obvious is that

[18] For instance, cycling of air conditioners so that they operate 50% of each peak hour would cut demand, and thus wholesale prices too, since air conditioning is a third of California's peak load in summer hours.

[19] This overstates the case since some utilities and national monopolies (e.g., Electricitè de France) offer tariffs that induce significant demand responses to prices or basic costs.

supplies are purchased or contracts negotiated based on ex ante judgments that are always erroneous when reviewed ex post for prudency. Midway in the crisis the California regulator allowed long-term contracts of certain kinds, but even in the face of insolvency the utilities avoided this option because they were still subject to prudency reviews that might result in profits from good decisions passing to customers and losses from bad decisions being borne by shareholders. This discrepancy between the processes of decision and review was avoided previously by exempting purchases in the spot markets of the PX and SO from review, but with the consequence that utilities had no options to hedge their risks.

Less obvious are the intricate ways that restructured markets affect the default-service obligation. For example, in competitive markets, energy and transmission are separate products, prices for transmission (and distribution) are spatially differentiated, and all prices vary greatly with time and events. Yet the standard means of universal service, inherited from the previous era, is a basic-service tariff that bundles all these together, erases most incentives for efficient usage in response to wholesale prices, and suffers from adverse selection in the customers served. The default provider has strong incentives to offer retail options to induce price-responsive demand behavior, but this runs counter to the regulator's imperative unless the entire issue is re-defined in terms of the unbundled ingredients of service provision, with separate accounting for cross-subsidies (as U.S. federal law requires for telecommunication services).

A summary of this section is that the financial collapse of the utilities in the California crisis showed the hazard of assigning to utilities all the risks of higher prices in wholesale markets. The insurance implicit in vertical integration and the regulatory compact ends when liberalized markets begin; the old risks remain but in the new regime the terms of trade between sellers and buyers are pecuniary risks for each party. As in other industries, these risks should be shared along the supply chain, and presumably the common interest of sellers and buyers in mutual insurance ensures they will be shared efficiently if the market rules do not restrict contracting. Each impediment, ranging from exclusion of hedging contracts to stifling of retail competition, increases the prospect that external shocks cause severe financial consequences. At the end of the chain, the regulator's relations with default providers are peculiarly sensitive to wider social concerns and the imperative of universal service. A new relationship is needed to make price-responsive demand a reality by encouraging utilities to offer innovative service options.[20]

Stanford Business School, Stanford CA 94305-5015 USA; rwilson@stanford.edu.

Manuscript received August, 2001; final revision received February, 2002.

[20] Among flaws revealed by the California crisis, most basic was that the retail sector was completely immune until months after the system collapsed. After rate increases were approved in May 2001 the crisis dissolved abruptly when peak demands during the usual June heat waves were far below previous years. Exhortations to reduce usage had some effect, but reductions in peak loads by over 10% reflected higher base rates and rates increasing steeply with usage.

APPENDIX

MITIGATION OF MARKET POWER

This appendix describes briefly some of the methods used to limit market power in the electricity industry.

A. *Ownership and Governance*

In most countries, market liberalization begins by privatizing state enterprises; in others, vertically integrated utility companies are divided into units functionally specialized in generation, transmission, distribution, or retail sales. This follows the scheme proposed by Joskow and Schmalensee (1983) who argue that, subject to sufficient safeguards against market power, the generation sector can operate via competitive markets if it is supported by a centrally managed transmission system organized on the principles of common carriage, and both are separated from regulated parts of distribution and retail sales. Continued regulation of the "wires" businesses of transmission and distribution is standard but in some jurisdictions operation of the transmission system is assigned to a state-franchised organization (the SO) charged with ensuring provision of the infrastructure of transmission and reserves required by efficient markets. In either case, the SO or transmission owner (Transco) inherits the software and technical personnel familiar with system operations. In the U.S., anxiety that a Transco could exert monopoly power or favor unregulated affiliates stems from the practices of gas pipeline companies, and the federal regulator has indicated its intention of separating ownership from transmission management.[21] But there is equal anxiety about whether the governance structure of an SO ensures efficient operations: the U.S. regulator dismissed the California SO's board representing stakeholders and demanded appointment of independent experts as in the Northeast systems, only to see the governor replace them with political appointees without prior experience in the industry, save one. No proposal for management of the SO as a franchise is sufficiently developed to ensure strong incentives for efficiency.

Governance arrangements arc potentially hazardous to entrants. Suppliers on the governing board of an SO can argue for technical requirements or compensation that amount to barriers to entry. These impediments might be allowed in systems organized as legal cartels as in New Zealand, but they are also possible where the governing board of the SO represents stakeholders. Then magnitude of the administrative hurdles is substantial: an entrant into New England must win approvals from twelve technical panels of the regional power pool association, and in California no general policy regarding new connections was approved during the first three years.

A peculiar feature of those systems derived from power pools are capacity payments, intended to attract new capacity and to retain obsolete plants that would otherwise be unprofitable to maintain. Requirements for installed capacity induce capacity payments when these obligations are tradable in auxiliary markets, as in the U.S. Northeast. A typical example of good intentions gone awry was Britain's capacity payment based on the product of an estimated probability of insufficient capacity, and an assigned value of unserved demand that was set administratively: Wolak and Patrick (1998) reports that the estimated probability of outages was ten times the actual frequency. Theory establishes that indeed a capacity payment is optimal if retail demand is inelastic and service cannot be curtailed, 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. 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

[21] In the U.S., pipelines' monopolistic practices include discrimination in terms and prices and withholding of capacity, except firm service at the maximum price allowed by regulators. 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 U.S. system allows a pipeline to lend one shipper's excess to another who is deficient and charge them both.

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 they prefer that the predicted load be represented inelastically in 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.

The relevance of the measures cited below 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 dominant incumbents. But this conclusion depends crucially on the availability of sufficient transmission capacity, which in turn depends on governance of the entity that decides on expansions of transmission capacity. An incumbent can argue that expanding import 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. Thus an incumbent supplier should not be able to veto expansions proposed by demanders, as for example in Ontario where an Electricity Board has authority to insist on construction of new transmission lines.

B. *Contracts and Incentives*

With regulation pervasive in other aspects, and some confidence that retail sales will be competitive if not regulated, concerns about market power focus on generation.[22] 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 requiring divestiture of generation assets avoided strict requirements sufficient to ensure vigorous competition, Britain being the prime example. The boldest have been states in the U.S., notably California where the three utilities divested all (non-nuclear) thermal generation, but even they allowed fleet sales in which large segments of generation capacity were sold to single buyers, in many cases the unregulated affiliate of a utility from another state. Thus California ended up with five firms controlling major shares of the state's thermal generation.

Apart from divestiture to promote competition, there are several ways of mitigating market power via contractual remedies.

(a) One way requires that reliability-must-run (RMR) plants, which have monopoly power because they are needed for local voltage support, must operate under long-term contracts with remuneration based on audited incremental costs. When reserve margins are thin, similar provisions could be applied to plants with unique capabilities to meet peak loads. Similarly, in Sweden the SO negotiates long-term options on capacity that can be used to alleviate transmission congestion. This seemingly simple approach can be difficult in practice: over a third of capacity in California was assigned to RMR contracts, and various strategic manipulations were persistent problems, in part because initially 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 way requires a firm with market power to be heavily hedged by long-term forward contracts for delivered energy at fixed prices. After such prior commitments, only the residual portion of the firm's output is affected by spot prices so its incentive to offer prices close to marginal cost is strengthened. Various versions are called legislated hedges, vesting contracts, or contract cover. In Britain initially and Australia still, the hedges purportedly worked well to sustain incentives for output until the contracts expired. In Alberta the percent hedged was so high that price variation was damped, and entrants were excluded by hedges contracted between a company's generation and distribution subsidiaries—in fact, hedging was so pervasive that prices in the spot market served as transfer prices between generation and distribution subsidiaries.

A sequence of markets has similar effects. Commitments made in earlier markets strengthen competition in later markets (Allaz and Vila (1993)).

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

(c) There are several ways to simulate the effects of hedges. The simplest requires each dominant firm to auction some percentage of its output, usually in the form of long-term contracts, as in Alberta and Texas recently. An important proviso is that 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 prevalent in early configurations of the New Zealand and American (except California) industries, and in Scandinavia where local distribution companies own substantial capacity.

In developing countries, liberalization often applies initially only to wholesale markets, while distribution companies retain monopoly franchises in retail markets; in such cases it suffices 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.

(d) Long-term relational contracts usually impose some limits on market power. New Zealand's Market Surveillance Committee has broad powers to implement efficiency-improving changes to the market rules and to sanction abuses of market power. However, similarly named committees elsewhere have authority only to monitor performance and to address occasional reports to regulatory agencies—and in Alberta the committee resigned in frustration.

C. *Re-Regulation*

The threat of re-regulation is explicit in New Zealand's unregulated system, and elsewhere it is implicit in the monitoring done by national commissions, as in Scandinavia. The salient example is California.

During the year of high wholesale prices in California, many observers argued that market power was the cause. Prices averaged somewhat higher in every other state in the West so it was difficult to distinguish between market power exercised by withholding capacity, and scarcity resulting from reduced imports of hydro power with resulting higher prices for the fuel and emission allowances needed by thermal generators—and ultimately, downtime for maintenance and repair. Nevertheless, a variety of interventions re-regulated aspects of the market. After the credit of the utilities was depleted, the state became the single-buyer of 40% of supplies (using general funds replenished by issuing bonds payable by a surcharge on retail sales), mostly from long-term contracts whose terms and prices were not disclosed on the grounds that secrecy impaired suppliers' market power; also, legislation established an official power authority to increase capacity. The SO restricted suppliers' flexibility of maintenance and scheduling. The federal regulator imposed substantial penalties on a participant completing less than 95% of its transactions before real-time, and directed that utilities could purchase directly from their affiliated generation companies. This was especially important because over half the state's capacity comprised these affiliates' hydro and nuclear generators that previously the state required to be offered in the power exchange's day-ahead market at a zero price. Attempts to restrict exports failed because they violated federal law, and indeed the federal regulator rejected many of the measures proposed. The two main interventions by the state and federal agencies aimed to cap prices.

(a) The SO imposed price caps, first \$750/MWh, then \$500, \$250, \$150. These had some effect but they jeopardized reliability. Suppliers increased exports to other states without price caps, and demanders waited until the SO's price-capped real-time market to purchase amounts sufficient to meet their loads, relying on the SO to procure supplies at the last minute and at any price by direct negotiation with out-of-state firms. (The utilities did, however, exploit their monopsony power by buying up must-run supplies offered day-ahead at zero prices in the power exchange.) This prompted the game of "mega-Watt laundering": supplies purchased by the SO at the last-minute might be exports sold in forward markets, so physically the power never left the state. One lesson learned was that a price cap is meaningless unless the SO curtails demand when supplies offered at the price cap are insufficient. The obvious fact that a price cap in one area of an interconnected region discourages

imports and promotes exports eventually led the federal regulator to establish a price cap for the entire West.

(b) Federal interventions began by dismissing the stakeholder Board of Governors of the SO, prohibiting the utilities from trading in the power exchange, and terminating key provisions in the tariff of the power exchange, which then ceased operation.[23] The regulator imposed a "soft" price cap requiring suppliers to cost-justify bids over $150, and later, out-of-state suppliers were required to export to California whenever reserve margins were low. In May of 2001 the regulator essentially re-regulated the energy market for the next 18 months by capping the price during any hour with low reserves at the highest of the generators' marginal costs, and in other hours at 15% less.

When a financial crisis occurs in an industry as critical as electricity the political response, if one judges from California, is vigorous re-regulation. California's retreat from liberalization, and its insistence that the crisis was caused by suppliers' abuses of market power, halted plans for wholesale power markets in other states and countries.

Some measures addressed deficiencies in the original legislation and implementation. The most important remedies accelerated construction of new plants that enabled over 10% new capacity within 18 months, allowed utilities with default-provider obligations to hedge their risks via purchases from affiliates and long-term contracts, and raised retail prices slightly in January and then significantly in May, quickly stimulating a substantial demand response that ended the immediate crisis. But the wholesale market had collapsed earlier, and because attention focused mainly on alleged abuses of market power by suppliers, California's liberalized markets were unlikely to resume without significant interventions by the state regulator and state agencies. Meanwhile, the federal regulator proceeded with plans to establish a regional transmission system operator for the entire West based on substantially liberalized markets, setting the stage for long struggles between federal and state authorities.

REFERENCES

ALLAZ, B., AND J.-L. VILA (1993): "Cournot Competition, Forward Markets and Efficiency," *Journal of Economic Theory*, 59, 1–16.

CHAO, H., AND S. PECK (1996): "A Market Mechanism for Electric Power Transmission," *Journal of Regulatory Economics*, 10, 25–59.

CHAO, H., S. PECK, S. OREN, AND R. WILSON (2000): "Flow-based Transmission Rights and Congestion Management," *Electricity Journal*, October, 38–58.

CHAO, H., AND R. WILSON (2002): "Incentive-Compatible Evaluation and Settlement Rules: Multi-Dimensional Auctions for Procurement of Ancillary Services in Power Markets," *Journal of Regulatory Economics*, to appear.

JOSKOW, P., AND E. KAHN (2001): "A Quantitative Analysis of Pricing Behavior in California's Wholesale Electricity Market During Summer 2000," Paper 8157, National Bureau of Economic Research, Cambridge MA.

JOSKOW, P., AND R. SCHMALENSEE (1983): *Markets for Power*. Cambridge MA: MIT Press.

JOSKOW, P., AND J. TIROLE (2000): "Transmission Rights and Market Power on Electric Power Networks," *RAND Journal of Economics*, 31, 450–487.

KREPS, D. (1987): "Three Essays On Capital Markets," *Es Revista Espanola de Economia*, 4, 111–145.

NEW ENGLAND SYSTEM OPERATOR (1999): "Information Advisory on the Operating Reserve Markets," August, Holyoke MA.

[23] The actions directed against the power exchange stemmed from conflict between federal and state regulators. When the federal regulator proposed to eliminate the requirement that utilities trade through the PX, the state regulator claimed the prerogative to continue the requirement, whereupon the federal regulator revoked the tariff provisions for mandatory trading to force state compliance. Although the revocation was to be effective several months later, and a new tariff could be filed, trading volume in the PX's markets quickly collapsed to levels insufficient to recoup its operating costs.

OFFER (1999): "Pool Price—A Consultation by OFFER," February. London UK: Office of Energy Regulation.

OFGEM (1999): "OFGEM Consultation on Rises in Pool Prices in July," September. London UK: Office of Gas and Energy Markets.

STOFT, S. (2002): *Power System Economics: Designing Markets for Electricity*. New York NY: Wiley-IEEE Press.

WOLAK, F., AND R. PATRICK (1998): "The Impact of Market Rules on the Price Determination Process in the England and Wales Electricity Market," Draft, Stanford University.

WILSON, R. (2001a): "Activity Rules for an Iterative Double Auction," in *Game Theory and Business Applications*, ed. by K. Chatterjee and W. Samuelson. Boston: Kluwer Academic Press, Ch. 12, pp. 371–386.

——— (2001b): "Ingredients of Liberalized Power Markets," Draft, Stanford Business School.