Part I
Market Reform Evolution

Chapter 1
Reevaluation of Vertical Integration and Unbundling in Restructured Electricity Markets

HUNG-PO CHAO[1], SHMUEL OREN[2], AND ROBERT WILSON[3]

[1] ISO New England, USA; [2] University of California, Berkeley, USA; [3] Stanford Business School, USA

Summary

This chapter critically reviews the argument for vertical integration in the electricity industry, and also the argument for restructuring based on unbundling of its products and organizations in favor of market mechanisms. The authors conclude that both arguments are deficient, and that a balanced mixture of vertical integration and liberalized markets is superior to the extremes. Their central conclusion is that efficient management of the risks inherent in the electricity industry requires that restructuring retain universal service for the core of non-industrial customers who rely on regulated rates smoothed over time to recover the costs of service.

1.1. Introduction

This chapter addresses basic economic issues posed by restructuring. The central issue is whether the overall technology of the industry – wholesale generation, transmission, and retail service – necessarily implies more or less vertical integration. It was long thought and is still being argued by many that vertical integration of retail utilities was essential for efficient investments and operations (e.g., see Michaels, 2006). On the other hand, restructuring has often been motivated by the view that the purported advantages of vertical integration are obsolete, that liberalized markets can work well, and that they bring stronger incentives that are likely to result in more efficient investments and operations (e.g., California Public Utilities Commission, 1993). The argument presented here is that neither view is conclusive – that pros and cons can be mustered on either side without any clear indication that one or the other extreme is better.

In prior work (Chao et al., 2006) the authors argued that restructuring of the electricity industry should develop along a middle path between the extremes of vertical integration and liberalization of wholesale and retail markets. This middle path establishes the boundaries of the firm – the extent to which a retail utility should retain some degree of vertical

integration. A key element of this choice is the make-or-buy decision about whether to own and manage supply resources, or to rely on wholesale markets via either spot purchases or longer-term contracts. A middle path also requires restructuring of regulatory policies and redefinition of the regulatory compact to recognize the effects of investment, purchasing, and contracting decisions by utilities in the context of liberalized wholesale markets, and to strengthen incentives for efficient operations and demand response. Moreover, the optimal extent of vertical integration is ultimately determined by the requirements for efficient allocation of risk bearing. After restructuring, the most important determinants of the optimal degree of vertical integration concern risk management, which affects the cost of capital – the ultimate measure of financial risk – and supply reliability and resource adequacy – the ultimate measures of physical risk. (See also Correljé and De Vries, Chapter 2 in this volume.)

From the perspective of risk management, the mutual interests of suppliers of generation and retail service enable risk sharing that mitigates financial risks. Depending on local circumstances, their shared interests imply a greater or lesser degree of reliance on markets and contracts, or on direct ownership that perpetuates some degree of vertical integration. For example, a utility might meet some resource adequacy requirements by contracts or by purchases in capacity markets, and also own generation facilities that serve its core retail customers within a regulatory scheme that continues the traditional regulatory compact, albeit with stronger incentives from market forces and performance-based regulation.

Section 1.2. begins by reviewing the case for vertical integration of utilities that prevailed through most of the twentieth century. Section 1.3. examines anew these arguments in the current context and finds them greatly altered – in part by the evident successes of some aspects of restructuring. The discussion of economic issues in Section 1.2. includes a summary of explanations of vertical integration in the literature. This discussion is necessary because ideas from this debate have greatly influenced restructuring decisions by regulators and legislators, especially in Europe recently. It also clarifies the distinction between financial and organizational "unbundling" of a utility's vertical components – wholesale generation, transmission, and retail service – and unbundling of the corresponding products. In the regulated era, the organization of the electricity industry stemmed from vertical integration of utilities in all respects, while in the past decade much reorganization aimed at segmenting utilities into their vertical components in conjunction with unbundling of their products. In several cases, organizational unbundling of firms was seen as a necessary or desirable complement to unbundling of products to facilitate liberalized wholesale markets. Although organizational disintegration was rejected in most other liberalized industries (transport, telecommunications, etc.), regulators and legislators favored dissolution of vertical organization in the electricity industry for reasons that are reviewed.

Section 1.4. reviews some of the unsolved problems of liberalized markets, including both those that cannot be solved efficiently by market processes and those that have not yet been solved adequately by market restructuring. Section 1.5. develops the case that risk management considerations are major determinants of the degree of vertical integration in terms of organization and ownership and vertical contracting. Section 1.6. concludes by outlining some implications for the evolution of restructuring. This discussion introduces scenarios in which a desirable degree of vertical integration coexists within liberalized wholesale markets for unbundled products, and which allow a utility to serve core customers at regulated rates while others opt to purchase from competing suppliers.

1.2. The Historical Motives for Vertical Integration

The origin of vertical integration in the electricity industry lies in a dominant public interest. Like other infrastructure industries – water, transport, communications – the energy industries were recognized as essential for economic development. Universal service, efficiently supplied at minimum cost, was imperative. In many countries these needs in the case of electricity were addressed by monopolies conducted or owned by local or national governments, and in some cases by government projects or subsidies; e.g., in the United States by the Tennessee Valley Authority, Bonneville Power Administration, Western Area Power Authorities, and the Rural Electrification Administration. The prevalence of government monopolies and government-sanctioned monopolies had three sources. One was technical, resulting from the advantages of alternating current synchronized over grids spanning large regions. Another was economic, resulting from the large scale of transmission and distribution (T&D) systems and the large scale of some generation facilities, especially hydroelectric dams but also the most efficient coal-fired and, later, nuclear plants. The third was financial, because the government was the sole or chief source or guarantor of sufficient capital at low cost. All three reflected the capital intensity of the technology used by the power industry, certainly in T&D, and in combination with fuel intensity in the case of generation. Historians of economic development view the twentieth century, in part, as an era of accumulation in which massive investments established the infrastructure on which a modern economy depends. (Chandler, 1969; Devine, 1983)

The industry's organization differed in those countries like the United States, Japan, and Germany that relied heavily on investor-owned utilities (IOUs). Although Nebraska and some municipalities developed public power systems, and federal projects were important elsewhere, within the United States most major urban areas depended on IOUs for provision of retail service. The role of IOUs stemmed from a conjunction of public and private interests. The public interest in universal service at minimum cost was matched by firms' interest in obtaining ample capital at low cost. The states established Public Utility Commissions (PUCs) to regulate the industry (except federal regulation of interstate trade), with authority to mandate the quality, conditions, and terms of retail service (Bonbright, 1961). In return, each utility obtained an exclusive regional franchise, except for municipal utilities and rural cooperatives, which were exempt. In principle, this was a retail monopoly but it evolved into a total franchise that encompassed local supply, transmission, and distribution as well as retail service. A state's grant of monopoly franchises on transmission and generation was artificial since it derived from comprehensive cost-of-service regulation rather than basic economic considerations. It was fundamentally at variance with federal legislation and regulation, but enforced by each state's control of siting of facilities, cost recovery from retail rates, and authority to exclude independent power producers (IPPs) from selling to retail customers.

Under the old "regulatory compact," risk management was provided through an insurance mechanism by vertical integration along the electricity supply chain. The single utility ownership of generation and transmission facilities buffered wholesale price volatility. Retail regulation smoothed the rate effects of cost changes on customers, imposed an obligation to serve, and offered utility shareholders a reasonable opportunity of recovering investments with a largely assured rate of return. Although all the risks – both physical and financial – were socialized to a high degree, customers bore the residual risk.

Importantly, a utility was assured full recovery of prudently incurred investments and expenses, including the cost of capital. This part of the regulatory compact was implemented by nearly level retail rates; that is, a utility's recovery of an approved cost (one accepted into the rate base) was amortized over many years, with repayments obtained

from retail revenues. The regulatory compact was a perfect means of obtaining capital from private sources to build a growing industry – because cost recovery was assured, utilities obtained capital from financial markets at low cost without drawing on public funds. Amortization of cost recovery reduced risks for lenders and shareholders, and equally, it reduced the volatility of rates paid by retail customers. For regulators, cost-of-service regulation brought difficulties judging prudency and measuring costs, and they were often dismayed by a utility's weakened incentives for cost minimization and strengthened incentives for capital-intensive projects (CPUC, 1993; Joskow, 1997). But until the last decade before restructuring these deficiencies were viewed as of second-order importance compared to the advantages.

A utility's monopoly on local generation and T&D was implemented by vertical integration of all aspects, including organization. The electricity industry has a linear supply chain from fuel to generation to transmission to distribution to service delivery. Each utility integrated backward from retail service to encompass at least generation, and occasionally some fuel sources. There were two motives for extension of a utility's monopoly backward into the supply chain, and with it the resulting vertical integration. One was the advantage of a single coherent investment strategy. Given the load-duration profile and the costs of building and operating generators, there is a particular mix of generation technologies that serves the load at least the overall cost in the long run. There is also an optimal configuration of the transmission grid and locations of generators, and moreover, an optimal substitution between local generation and transmission to access distant generation – as well as occasional use of local generation to alleviate congestion on transmission elements, sustain voltage, etc. The second motive was the advantage of consolidated operations. Centralized dispatch of generation and transmission had the explicit objective of minimizing the total cost of serving the load subject to constraints intended to ensure service reliability and protect the transmission grid from cascading failures.

1.2.1. Theoretical framework

These motives were always based on ideal realization of the alleged advantages in investments and operations. In reality the actual results were often driven by practical financial considerations, as explained below. Even so, a substantial body of economic theory was constructed to explain the prevalence of vertically integrated utilities (e.g., Williamson, 1975, 1985). Its main ingredients were as given below:

- *Public good*. The T&D system is the enabling infrastructure of the power industry. Tight control, operating on very short time frames, is required to sustain service reliability and to avert cascading failures of grid elements and generation units. Also necessary are uniform standards and procedures among interconnecting segments of the grid.
- *Natural monopoly*. Duplication of T&D facilities is wasteful except where it improves grid security or service reliability.
- *Economies of scale*. Natural monopoly was extended to generation by citing the large size and capital requirements of efficiently scaled units and plants. This argument applied mainly to hydro projects and base-load plants using coal and later nuclear fuels.
- *Economies of scope*. This catch-all category (in principle, a subset of economies of scale) cites advantages from tight coordination, such as the above-cited advantages

of centralized investment and operations. It also includes advantages from substitution (e.g., generation capacity usable for either energy production or a contingency reserve, and generation used to alleviate transmission congestion), and the possibility that standards, technology, information systems, and skills used for one kind of generation are applicable to other kinds and to engineering control of the grid. Savings in metering, billing, and financial settlements are sometimes included in this category.

- *Economies of transaction costs*. Despite its name, this category refers not to costs of metering and billing, but to difficulties and risks in contracting. Its premises include asset specificity and incompleteness of contracts. A seller's investment in a transmission or generation facility is irreversible and long lived, and the facility cannot be moved or used for another purpose. The value of the investment is therefore tied specifically to expected use by or sale of output to buyers. If there is a single buyer then an initial contract between them might seem to ensure that the seller obtains the value he anticipates when he commits to investment and construction. But a contract that covers all contingencies is usually infeasible, and in those unlikely contingencies that are not covered (or if the buyer can renege) the seller might not be able to renegotiate with the buyer to recover the sunk costs of investment. Anticipating this, the seller might not undertake the investment initially. This scenario is the basis for the argument that contracts may be insufficient to stimulate adequate investments, and therefore vertical integration of the seller and the buyer might be necessary to ensure that efficient investments are undertaken.

These technical explanations of vertical integration did not, however, address the more practical aspects that were constantly at the forefront of regulatory considerations. These were dominated by financial considerations that are described next.

1.2.2. Financial motives

In keeping with their primary responsibilities, PUCs focused on capturing the advantages of vertical integration for retail service. Cost-of-service regulation was the means to obtain mandated universal service at minimum cost and with high reliability. In some ways a vertically integrated utility was easier to monitor, its total costs were easier to measure, and it could be held directly accountable for deficiencies of quality or reliability. Universal service required subsidies to those residential and commercial customers who were more expensive to serve, and vertical integration offered the expedient of relying on implicit cross-subsidies rather than explicit financial subsidies. Although industrial customers were especially disadvantaged by this policy, the inefficiencies of cross-subsidies were secondary to the political influence of residential and commercial customers. The magnitude of cross-subsidies declined in later years as the development of efficient plants of small size and the growth of co-generation enabled an industrial customer to negotiate lower rates, since it had the option to self-generate to serve its own load.

Most important for PUCs was that retail rates could be smoothed over time by amortizing the utility's recovery of its costs, and the cost of capital could be minimized by indirectly invoking the credit of the state. Full recovery of costs via retail rates necessarily implies that retail customers ultimately bear nearly all of the financial risks; indeed, this is the first fundamental principle of the regulatory compact. For many customers their aversion to volatile rates is profound, in part because on short timescales they have limited options to alter usage patterns or to invest in alternative appliances and production technologies, and generally they cannot obtain financial hedges against fluctuating rates.

However, the second fundamental principle is that cost recovery is amortized so that rates can be smoothed inter-temporally over long periods.

The feasibility of this scheme stems basically from the difference between the high volatility of fuel and power prices in the short term and their low volatility over the long term. Short-term volatility is mainly cyclical and load-based, ranging from daily variation to seasonal weather cycles to business cycles. Thus, retail customers are exposed only to secular risks and trends, and only gradually. These include trends in fuel prices and generation technologies, and inevitable mistakes such as misestimates of the amount and location of load growth. However, this rosy scenario was upset in the years after the oil embargo of 1973 and before restructuring in the 1990s by cost overruns for nuclear plants and by the prices guaranteed to "qualifying facilities" (QFs) as specified in the Public Utilities Regulatory Policies Act (PURPA) of 1978. But until well into the 1980s cost-of-service regulation was generally viewed as successful in spite of the difficulties during the 1970s and 1980s from gyrating fuel prices, monetary inflation and high interest rates, and technical advances that rendered major investments inefficient.

The natural monopoly aspects of T&D systems implied regulation and control of rates. Most investments in T&D facilities could not be recovered by marginal-cost pricing or by congestion pricing. Therefore, cost-of-service regulation that provided recovery of investment and maintenance costs extended naturally to T&D. In later years there were instances of performance-based regulation (Hunt, 2002), and, rarely, of merchant transmission investments in direct current lines, but overall the expansion of the grid was a massive investment in infrastructure that continued until in the United States and Canada it is now composed of only two interconnected systems plus one within Texas. Although essentially a public asset, the grid is largely privately owned by utilities and financed mainly by recovering the costs from charges included in the rates paid by retail customers. Like some other infrastructure networks (railroads, telecommunications, gas pipelines) it was regulated according to principles of contract carriage – until superseded in 1996 by common carriage when the Federal Energy Regulatory Commission (FERC) required open access and nondiscriminatory pricing. Reliance on private ownership and contracting prevailed in some countries (e.g., Germany and Japan) while state-owned transmission companies developed the grid in others (e.g., the United Kingdom, France, New Zealand, and Scandinavia). The latter developed systematically but those relying on local utility-owned transmission developed through increasing interconnections among them as energy trading increased and the utilities increasingly relied on exchange agreements to improve reliability.

The financial aspects of generation were fundamental motives for vertical integration. Mentioned previously were the role of capital intensity, the scale economies of base-load plants before smaller gas-fired plants were developed in the 1980s, and possible economies of scope. These are reinforced by the great variation of loads on short time frames and the resulting high volatility of prices in spot markets, plus longer-term secular trends. Since a generation plant has a lifetime of 20 to 40 years, its inherent value is largely unaffected by short-term price volatility. Moreover, the supplier and a buyer such as a utility have mutual interests to ensure each other against price variations, since every price that is good for one is bad for other. Thus, one might surmise that long-term contracting will ensure investments in generation capacity; indeed, the investor can use the contract as security to obtain loans to finance construction.

This scenario is jeopardized, however, by two factors. One is that a 20- or 40-year contract differs from utility ownership of a plant only in its lack of direct investment and operational management and control, but requires comparable justification to the

PUC that it is prudent, and further that it is invulnerable to the supplier's default or bankruptcy. The other factor is that the investor or its lender prefers a contract that fixes both price and quantity, while the utility prefers flexible dispatch to meet changing loads and overall demand growth, so their mutual interest in price insurance is diminished by their opposing interests in "volumetric" insurance. All the intermediate contract forms (e.g., option contracts, tolling contracts) require at least one and usually both parties to bear risks of one kind or another. For the investor or its lender, risk reduces the value of the investment, and for the utility, any residual risk borne by its shareholders is inferior to assured cost recovery if it undertakes the investment itself – in the usual circumstance that it can obtain capital at lower cost than non-utility investors – and thereby transfers the risk to ratepayers.

These financial considerations in generation investments are variants of the factors invoked in the analysis of economies of transaction costs described earlier. A contract sufficiently complete to deal with all contingencies is too complex to be practical, and even if it were feasible and could miraculously insure both parties so that for the utility it is not inferior to inclusion of the new plant in the rate base over the lifetime of the plant, approval of the contract in a prudency review would be problematic – and in some contingencies might not be enforceable. For example, if the contract extended over decades then during a prolonged period of low prices the supplier might default on the contract.[1] The realistic contracts are therefore short-term (usually a few years, rarely 10) and incomplete, with both parties bearing shares of the price and volumetric risks. Until shortly before restructuring, this picture doomed non-utility generation from the start. With its low cost of capital and other advantages, a utility could always undertake a more efficient portfolio of investments than could private investors exposed to price and/or volumetric risks, and do so without sacrificing dispatch control.

In the United States, PURPA first forced utilities to purchase generation from the QFs. This wedge initiated further opening of wholesale markets. Next were the Energy Policy Act of 1992 and the FERC's ensuing Orders 886 and 888 that forced utilities to allow open access to unused transmission at nondiscriminatory prices. This wedge also initiated bilateral contracting between non-utility power generators and large industrial customers in the late 1990s.

1.2.3. The hidden assumptions

The justifications of vertically integrated utilities contain hidden assumptions. As described in Section 1.3., many of these were revealed by actual experience after restructuring. Several that bear on how one interprets the foregoing arguments for vertical integration are listed below.

- *A utility obtains capital at lower cost than its supplier*. This assumption might seem to contradict theories of finance, but in fact it was realistic before restructuring. The seeming contradiction stems from the fact that one can invest equally in shares of the contracting parties (the independent generator and the utility). It therefore seems that an investor can hedge against price and volumetric risks that affect the seller and buyer oppositely, and therefore the seller and buyer should obtain

[1] This assumes omission of provisions in recent contracts that enable the utility to take over the plant immediately in the event of default. Since restructuring, utilities' contracts with suppliers have increasingly included stringent provisions to protect against the consequences of default.

capital at comparable costs. In fact, however, cost-of-service regulation transfers the utility's risks to ratepayers, and an investor cannot easily invest in shares of the general ratepayer population, especially over the extended period of cost recovery. If the transfer were not mixed with cross-subsidies then an investor might hedge by buying shares of a generator and an industrial customer, but this strategy was ineffective in the regulated era. Thus, one must view utilities' lower cost of capital as both a cause and a consequence of vertical integration under cost-of-service regulation.

- *The utility is the sole buyer of generation*. Those arguments that invoke economies of transaction costs depend heavily on the bilateral character of supply contracts and on the dominant monopsony power of the local utility, derived from its exclusive retail and transmission franchises. The advent of bilateral trading of bulk power supplies over wide regions, enabled by open access to transmission on nondiscriminatory terms, undercut this argument, and it evaporated with the arrival of multilateral trading in power exchanges managed by system operators. Both parties to contracts now have comparable options outside their bilateral relationship, namely to sell or buy at prevailing spot prices or to contract on a longer-term basis with competing alternatives. The prospect that a generation investment would be hostage to the sole utility buyer of energy, and to its transmission system, must now be seen as a consequence of the utility's exclusive franchises, and thus an artifact of regulatory policy.
- *The owner of the transmission grid has the authority to manage it*. Another artifact of regulatory policy was a utility's right to exclude access to other parties, even if ample capacity were available after serving its native load. Arguments were mustered that the utility needed exclusive control to assure reliability and security, but with the advent of regional system operators (in 1998 in the United States for those systems not previously organized as power pools, but years earlier elsewhere) this argument dissolved. Indeed, the efficiency gains from regional dispatch and transmission allocation are now widely recognized. The argument for economies of scope also dissolved since nearly all components of this category now pertain to the system operator. What remains is the vestigial argument that coordinated planning of transmission expansion and generation investments might be more efficient if it were centralized, but even so the advantages occur at the level of regional systems rather than utility service territories.
- *Incentive effects are minor compared to economies of scale and scope*. The engineering expertise of the utilities was always admired, and the PUCs took some credit for the fine performance by attributing it to the high standards they imposed (one day of outage in 10 years was the prevailing standard), as well as the advantages of vertical integration in establishing operating standards and procedures and in internalizing the public-good aspects of system security. This credit was less convincing after engineering operations and many skilled personnel were transferred to the FERC-regulated regional system operator. The regulated era was always beset by complaints that the commercial parts of utilities were complacent bureaucracies, and motivated more by cost recovery at inflated costs of capital than by minimizing overall costs (e.g., by favoring capital-intensive projects).

In later years these complaints led PUCs to experiment with performance-based regulation, rate caps, negotiated rates for industrial customers, and other devices. Behind these complaints, however, lay the basic fact that a utility's financial incentive was muted by assured cost recovery and thus by full insurance against contingencies – and after the PUC

accepted an asset into the rate base, by insurance against errors of judgment. The source of this comprehensive insurance was the regulatory compact. In some states, guaranteed cost recovery escalated rates to levels higher than neighboring states that then attracted away industrial and commercial firms seeking lower energy costs, or required negotiated industrial rates to forestall self-generation. In the United States restructuring was precipitated by one such state when the California PUC announced in 1993 that it would consider new regulatory principles and policies based on greater reliance on markets (see Borenstein et al., 2002). A chief consideration was the view that generation investments would be more efficient if private investors rather than ratepayers were to bear the consequences of erroneous judgments. Since insurance mutes incentives, one of the alternative policies it outlined, the one adopted in 1994, withdrew some or all of the provisions for assured recovery of investments in generation.

These four assumptions, mostly hidden in the standard justifications and explanations of vertical integration, are typical of a longer list. They are emphasized here because they exemplify a tendency for the merits to be addressed within the regulatory policies, institutional structure, and market rules that sustain vertical integration. Section 1.3. takes the opposite approach and describes the fundamental changes that preceded and followed restructuring. Restructuring introduced new regulatory policies and market structures that reflected a new view that vertical integration is not intrinsic to the electricity industry; indeed, one can assemble a comparable argument that electricity is amenable to an industrial organization that relies heavily on liberalized markets.

1.3. The Case Now for Liberalized Markets

The case for liberalized wholesale markets is now examined from several perspectives. First we provide a brief review of the situation after restructuring. Then theoretical arguments are examined in light of empirical evidence. In both cases the discussion includes some developments in the two decades before restructuring in the United States, along with the experience after restructuring. The analysis focuses on those aspects that indicate the future role of vertically integrated utilities. The discussion is organized around four utility functions after restructuring: system operations, wholesale markets, retail service, and generation.

1.3.1. System operations after restructuring

Some aspects of the industry remain unaffected by restructuring. The importance of universal service is reinforced now because economic development depends on technologies that rely on reliable power supplies. The traditional role of lighting is now supplemented by digital information and communication systems. Heating and cooling applications that were considered secondary are now considered essential for much commercial activity. The increased role of reliability enhances the public-good character of the transmission system, which remains a natural monopoly. But management of the grid is now viewed as a technical task, one that the engineering profession is well able to conduct, and that can meet the highest standards when its span is regional. This consensus developed early in those systems with national transmission companies, but in the United States it evolved from cooperative power pools and from the observation that local utilities were well able to integrate supplies purchased from QFs into their routine operations and dispatch procedures.

Control areas confined to utility service territories are an impediment to coordination over the wide areas now required. The magnitude of coordination problems at seams will eventually reveal the optimal span of regional operations, but only the largest utilities are viable candidates for retaining their own control areas. Equally, a utility's motive to hoard its transmission facilities to serve its native load impairs efficient allocation of generation and transmission capacity. Recognizing this, orders by American regulators have steadily mandated or encouraged formation of independent system operators (ISOs) and regional transmission organizations (RTOs) with authority to manage regional transmission systems on a daily basis, and responsibility for ensuring open access and nondiscriminatory pricing. These changes are partly organizational, but also financial since they invoke principles of common carriage, remove "pancaking" of transmission charges, and impose charges to recover costs of ancillary services and re-dispatch to alleviate transmission congestion.

1.3.1.1. Consolidation of grid and market control

From an economic viewpoint, the organizational specialization represented by system operators is notable for two features. Most important is that the externalities inherent in grid operations are handled by engineering procedures that enforce standards for reliability and security. The most obvious externality stems from the grid's role as a public good enabling transmission among locations. Because the grid is vulnerable to cascading failures, automatic switches open lines and disconnect generators to minimize damage to facilities. The consequences for suppliers and retail customers are partly transitory due to loss of power production and consumption, but their equipment and appliances can also be injured, and industrial customers can lose goods in intermediate stages of production and incur the costs of idle labor.

A subtler externality stems from differences between the technologies of supply and demand. On the supply side, sufficient fast-response reserves must be available to meet most contingencies because the ramp rates of generators are limited. The high value of ramp rate on the supply side has no counterpart on the demand side, since customers care only about whether power is on or off.[2] If there is only one customer (e.g., a utility) then it sees clearly that continuous power availability depends on its provision and payment for reserves. But if there are many customers then each knows that what it pays for reserves has little effect on its own access to power. This is a classic free-rider problem (if there are many customers then each prefers that others provide and pay for reliability) since the marginal effect of any one customer's contribution to reserves has a small effect on overall system reliability and therefore a small effect on the reliability of that customer's supply. The free-rider problem is significant even when the customers are a few utilities, since each sees that its own marginal value from a marginal dollar expended for reserves depends heavily on how much others contribute.

[2] A mathematical model leading to the conclusion that a completely decentralized market for energy and reserve capacity does not obtain full efficiency is by Chen et al. (2004). They conclude: "The decentralized market tends to depress ancillary service prices, which leads to the failure of the second welfare theorem. At each time the two market prices represent the market value of the on-line capacity of ancillary and primary services. From the system operator's viewpoint, the ancillary service is more valuable because of its higher ramping rate. However, from the utility's viewpoint, both services are identical, as long as they are available. In the decentralized market, the utility does not consider the ramping rate, since this is a constraint on production rather than consumption. As a result, the utility does not want to pay a higher price for the ancillary service."

Thus, without regulatory intervention and engineering command-and-control, markets for reserves are bound to be inefficient or even to collapse. These considerations extend beyond daily operating reserves to the general problem of ensuring adequate supply resources, including both generation and transmission capacity. To an investor, constructing a plant that will be idle most of the time seems a waste because it is called only rarely to meet contingencies, and this is equally true of a transmission line constructed to provide a backup for others that might fail. Thus, some form of payment to idle capacity is necessary to ensure that investments are sufficient. The practical issues of implementing resource adequacy requirements are addressed in Chao et al. (2006).

In principle, each customer could be charged for the *rate* of variation in its load, and thus pay for reserve generators with high ramp rates. However, a basic economic advantage of a regional electricity system is that variations of customers' loads largely cancel out, and aggregate loads are substantially predictable – the day-ahead prediction of the aggregate peak load in an hour is usually considered to be accurate within 3% or 4%. For this reason, retail pricing has largely ignored the possibility of charging for load variation other than on the basis of a customer's load-duration profile over an extended period such as a year and to some extent, by real-time energy prices revised every few minutes.

One could charge for load variation in some conditions of aggregate variation, such as the morning ramp at the beginning of the workday, or for increasing use of air conditioning when temperatures escalate on a hot summer afternoon, but in fact metering and pricing have not been developed to this degree of refinement.[3] Instead, system operations provide a buffer to ensure the steady matching of supply and demand, re-dispatching online generators and calling on reserves as necessary to follow the aggregate load. The buffer is partly automatic, since generators equipped with governors and automatic controls (AGC) adjust power output in response to frequency variations detected by sensors.

The automatic buffer provides an operator with an interval, usually considered to be about 10 minutes, in which to re-dispatch and call on reserves, beginning with hydro and spinning reserves, and then non-spinning reserves that require start-up and synchronization. The basic economic significance of system operations is that they supplant price-mediated market mechanisms in favor of command-and-control to ensure reliability and real-time matching of supply and demand. Market processes might conceivably be used to balance supply and demand almost continuously, but the costs and risks are too extreme to make them feasible on the short time frame that is relevant in a power system.

Thus, the economies of scope invoked to justify vertically integrated utilities in the United States are now mostly obtained by consolidation of grid management and wholesale spot markets in system operators. These developments in the United States imitated earlier initiatives in other countries with national transmission companies. Linking of the Scandinavian state-owned systems into the coordinated multinational NordPool was one model, and the other was the England–Wales Pool that began in 1989 [e.g. Hunt (2002) and its analogs in Argentina, Alberta, and Australia – followed later by New Zealand, Spain, and others, e.g. EPRI 2002 and Rudnick et al. (2005) and Barker et al. (1997)]. Some consolidated systems encountered initial operational and economic problems, and all were reformed later in some ways, but the basic principle that regional systems are more efficient and more reliable has not been challenged.

[3] However, after the California crisis the state provided $35 million to fund installation of interval meters at all large industrial and commercial customers, accounting for about a third of the aggregate load. Estimates of the costs of meters and the resulting benefits typically imply substantial net benefits; (cf. Borenstein 2004).

1.3.1.2. Unbundling of products

A second feature of system operations is unbundling of wholesale supply into constituent products such as energy, reserves, and transmission, which nevertheless are allocated jointly by a consolidated spot market conducted by the system operator. Unbundling of supplies into relatively homogeneous products like energy, reserves, and transmission recognizes the basic scarce resources managed by engineers. These products must be differentiated by attributes like time and location, and operational constraints like a generator's start-up time, minimum energy output, and ramp rate. Earlier arguments that this complex mix of products and attributes could not be efficiently priced are now, after successful implementations of time- and location-differentiated nodal pricing, confined to the operational constraints that involve nonconvexities, such as the start-up costs of a generator. Similarly, the various categories of regulation and operating reserves (spin, non-spin, replacement) are now priced systematically by recognizing that speed of response is the scarce resource, and therefore prices for slower reserves are limited by the feasibility of substitution with faster reserves.

Unbundling wholesale supply into standard products enables efficient allocation among multiple parties, and more specifically, unbundling facilitates markets and settlement procedures for multilateral trade. Markets for medium-term bilateral contracts for energy (e.g., Eltermin in NordPool and the UK Power Exchange) also rely on standard product specifications.[4] The responsibilities of a system operator now exemplify the economies of scope argument, since the engineers protect reliability and system security while also facilitating and/or conducting wholesale markets, and, indeed, procure resources needed for grid management from these same markets. For instance, the real-time balancing market is a market for both buyers and sellers, and also provides the bids from which the operating engineers obtain resources to follow the load, alleviate transmission congestion, and sustain voltage.

Even so, spot markets have been organized quite differently among various system operators. Some reflect technology, as in the first version of NordPool where ample hydro resources allowed self-scheduling and an emphasis on energy trading, and zonal pricing sufficed since the chief transmission bottleneck was between Norway and Sweden. Those adopting the England–Wales Pool model focused on efficient day-ahead scheduling of thermal generators using a comprehensive optimization of dispatch. The ISOs in the US northeast also adopted this model since it simply extended the procedures of pre-existing power pools. Their relative success during and after the California crisis in 2000–01 motivated FERC to propose a Standard Market Design (SMD) that is now adopted also by California, among others (FERC, 2002). The SMD's emphasis on comprehensive day-ahead optimization of all aspects does more than consolidate spot markets for products like energy, reserves, and transmission, since the optimization allocates available generation and transmission capacity by including unit commitments and scheduling along with assignments to energy generation and reserve status. Settlements are based on hourly locational prices for energy, reserves, and transmission, but the scarce resources are capacities rather than flows.

[4] A peculiarity of bilateral contracting is that the resulting demands for transmission need not result in an efficient allocation. That is, with bilateral contracting the use value of transmission depends on the pairings of sellers and buyers, whereas multilateral trading can provide the maximum value obtainable among all pairings. This potential problem has been insignificant in systems that have enough multilateral trading (e.g., 30–40% is scheduled centrally in PJM) to ensure that transmission is accurately priced to reflect scarcity values at the margin.

California considered the pool model but opted for an initial design closer to NordPool, including self-scheduling of generation and loads, a power exchange separate from the ISO, zonal pricing of transmission, and simple market clearing for standard products rather than comprehensive optimization (cf. Sweeney, 2006). This design was plagued by loose coordination and gaming of market rules almost from inception, and then virtually collapsed in a series of events initiated by scarcity of imported supplies from hydro sources.[5] In the United States it is now widely accepted that the system operator must retain tight control, and in particular, that its authority must extend beyond reliability to maintenance of orderly wholesale markets. For instance, FERC now allows ISOs to impose various protective measures on generators: must-offer obligations, bid caps, automatic procedures for mitigation of market power, obligations to respond to dispatch instructions, mandatory scheduling of maintenance, penalties for large deviations from day-ahead schedules, and a dozen more interventions that might be listed. Analogous problems occurred elsewhere (e.g., New Zealand) when supplies were scarce, but there is no exact parallel to the California crisis.

The panoply of ad hoc protective measures now imposed in the United States have not been adopted widely because other countries provide system operators with ample authority to ensure reliability and more discretion in managing their markets. The United States is unique in requiring an ISO to adhere rigorously to the terms of its FERC-approved tariff and market rules, and avoid any influence on energy markets. The California ISO was allowed the least discretion and was least able to bring its markets under control, but inability to weather a crisis is inherent in the strictures placed on all ISOs. The exact opposite can be seen in the United Kingdom's reformed New Electricity Trading Arrangements (NETA) system, where the transmission system is owned and operated by an independent transmission company (ITC), the National Grid Company (NGC), rather than a non-profit bureaucratic ISO. NGC must adhere to a Balancing Code for settlements but, within a scheme of performance-based regulation that rewards reductions in its grid management charge, it has wide discretion to manage the grid, including taking positions in the energy market to acquire reserves and counter market power. The NETA system is also the opposite of FERC's SMD, since all energy trading is conducted through private power exchanges for bilateral contracts, and physical feasibility is established hours-ahead rather than day-ahead (Newbery, 2006).

1.3.1.3. Lessons learned

The lessons learned from the recent experience with system operations can be summarized thus: The organization, governance, and procedures of a regional system operator are very important and very complicated. Creating such an entity is a major task from an engineering viewpoint, and the design of its markets is equally challenging from an economic viewpoint. The complexity of system operations ensures that an initial design must be revised as deficiencies are discovered. The task is worthwhile because the regional scope can enhance reliability and improve overall efficiency of the short-term allocation of generation and transmission capacity. The vigor and growth of wholesale energy markets attests to gains from trade from system operations and markets on a regional scale. On the other hand, the California crisis was a salutary warning that wholesale power markets are

[5] The origins and history of the California crisis are described by Blumstein et al. (2002) and Wilson (2002). An empirical analysis of the role of market power during the crisis is by Borenstein et al. (2002).

fragile when supplies are scarce, demand is inelastic, and utilities are obligated to serve but financially exposed. It is necessary to address these vulnerabilities – in terms of both resource adequacy and financial exposure – as argued later, but a backup remedy should be adequate authority and discretion for the system operator to intervene to stabilize its markets. That is, it should not rely solely on engineering procedures and established market rules when active intervention can stabilize markets or suppress the influence on market prices of dominant suppliers, and thus its scope should include some provisions for active management of markets.

There is no indication yet that any one design is best. The diversity of designs now working reasonably well in various countries and within the United States suggests that local factors are important. A principal reason that local considerations can be determinative is that engineering management of a transmission system is so well developed (and virtually uniform worldwide) that it assures most of the gains from a regional system regardless of which among several alternative market designs are used. Wholesale markets for energy might be bilateral or multilateral, decentralized or optimized, provided system operators have adequate means to ensure reliability and allocation efficiency. A relevant comparison observes that the United Kingdom relies on private markets for bilateral contracts, Australia relies on an energy-only real-time market, and NordPool uses a day-ahead and real-time market; and zonal pricing of transmission congestion suffices in the latter two. Seemingly quite different are the day-ahead and real-time markets in the Pennsylvania–New Jersey–Maryland (PJM), New York, and New England systems that include unit commitment and scheduling and co-optimization of energy and reserves, and insist on the importance of nodal pricing of congestion. Yet differences in performance among these systems must be considered of second-order importance compared to their overall successes.

This in particular requires that markets are workably competitive; that is, incentives promote productive and allocative efficiency. The California experience especially demonstrates the need for a design that discourages gaming of market rules. Gaming is essentially always due to a market imperfection, usually an unpriced scarce resource, and therefore a signal that efficiency can be improved by the measures that also eliminate gaming. For instance, the costly "dec game" in California was possible because congestion charges were imposed only between large zones and only day-ahead, which enabled those who caused intra-zonal congestion to escape congestion charges day-ahead and then be paid in real-time for alleviating the congestion they caused. The fact that nodal pricing eliminates the dec game illustrates the more general principle that even though many market designs are possible it is still true that efficiency requires that all scarce resources are priced, in this case intra-zonal transmission capacity.

1.3.2. Wholesale markets after restructuring

This subsection outlines the economic argument that a liberalized wholesale power market is potentially an efficient means of allocation among buyers and sellers of energy. This argument assumes, of course, that engineering aspects are conducted by a system operator whose procedures supplant market processes on the short time frames relevant for protecting reliability and ensuring continuous matching of supply and demand. The discussion focuses first on the spot market for multilateral trading, then extends the analysis to the forward market for bilateral contracting, and then examines the incentives for efficient investments. The argument is mainly theoretical but mentions indications that its

predictions are confirmed in some markets now in operation. The complexity of the markets now conducted by ISOs is evidence that implementation is difficult, but the concern here is whether wholesale markets are efficient in principle.

1.3.2.1. Models of restructuring

The pace and scope of liberalized wholesale markets differ greatly among countries. The focus below is on those with comprehensive markets, some of which extended their scope gradually (e.g., NordPool, and Australia, which evolved from Victoria's VicPool) while others liberalized in a single decisive act (e.g., England–Wales in 1989). There are two basic models: In one model the utility remains the single buyer, but regulators require the utility's "make or buy" decision to consider competing offers from IPPs. In the United States this approach took an extreme form due to the 1978 PURPA that essentially required a utility to pay its avoided cost for supplies from small plants (less than 80 MW) that used co-generation of heat and power or renewable sources of energy. The second model allows a market for bilateral contracting between IPPs and large industrial customers, augmented by provisions for enhanced opportunities for trading among utilities. This model was in effect in the United States in the period after federal regulators required that utilities provide open access to transmission on nondiscriminatory terms, and it too stimulated substantial investments by IPPs. It is currently seen in the interim phase of the European Union's directives for partial liberalization, and more specifically in the organization of the electricity industry in Germany.

It may be that most of the gains and fewer problems are obtained with these intermediate forms of liberalization. Indeed, even in the United States those restructured systems that have allowed utilities to remain substantially integrated are cited as more successful – and the complete divestiture of gas- and oil-fired plants by California's utilities is cited as one source of the crisis there. However, the aim here is to examine the viability of fully liberalized wholesale markets, and therefore to focus on those systems with comprehensive markets.

The argument for restructuring depends crucially on its most important innovation, which is management of the regional transmission system by a system operator such as an ISO (O'Neill et al., 2006). Initially the discussion simply assumes that a system operator manages the transmission grid, although some of the difficulties encountered by a system operator are mentioned in passing. This enables bypassing the public-good and natural-monopoly aspects of transmission, and some of the operational tasks required to assure reliability and system security. At the end of this subsection are comments on deficiencies due to inefficient allocation of risks, and an outline of the role of regulated utilities in improving risk management.

The main motive for a market is to realize gains from trade. This motive originates in some separation of ownership. Power markets (and cooperative power pools) began with exchange agreements among utilities to enhance reliability, extended to long-term trading of energy supplies based on differing costs and asynchronous loads, and then expanded to daily "economy" trading to minimize generation costs. The gains from trade in a modern market stem from its regional scope and from substantial separation of ownership between sellers (generators) and buyers (utilities and other load-serving entities, LSEs). Separation of ownership between generators and LSEs may reflect advantages from specialization, but restructuring has proceeded more on the premise that the main advantages come from stronger incentives.

One incentive effect is supposed to be more efficient investments when investors in generation bear the consequences of their decisions and operating decisions, rather than

relying on utilities' assured recovery of costs. The second incentive effect is vigorous competition among generators when they are sufficiently numerous and small to have little influence on prices. Competition among generators is imperfect in most wholesale markets, and therefore typically requires some regulatory interventions; e.g., price or bid caps and must-offer obligations are typical. The vigor of competition depends ultimately on sufficient investments in generation and transmission capacities.

At a minimum this requires that no firm is "pivotal" in the sense that its capacity is needed to meet a peak load in a transmission-constrained area. Over a longer time frame, competition stems from contestability, in the sense that incumbents cannot maintain high prices without stimulating new investments by entrants. Contestability has become a more effective constraint as smaller combined-cycle units can be installed in a few years, and for peak loads and offline fast-response reserves, combustion turbines (CTs) can be installed in a few months. Competition among LSEs is usually considered to be of secondary importance because their service obligations and the price inelasticity of retail demands limit their opportunities for strategic behavior to arbitrage between forward and spot markets.

1.3.2.2. Spot markets

The spot markets conducted by ISOs are multilateral; that is, they allocate supplies offered by several sellers to several buyers. They are also "smart markets" in that energy trades are optimized subject to constraints on transmission capacities and generators' ramp rates, required quality attributes (frequency, voltage), and operational procedures that ensure sufficient reserves to meet contingencies and avert cascading failures of equipment. A multilateral market is feasible only if several basic requirements are met. These include standardized commodities and qualities, accurate metering, explicit market rules, and settlement procedures that include assured creditworthiness of market participants. These requirements and other enabling aspects are now routine among system operators. Participants must subscribe to a contractual agreement that imposes reciprocal obligations, such as compliance with dispatch instructions, scheduling of outages for maintenance. After allegations of manipulations during the California crisis in 2000–01, the importance of a rigorous code of conduct and steady scrutiny by an independent market monitor is now universally accepted. Here it is taken as given that the mainly smooth operations of spot markets conducted by system operators are evidence that the basic enabling requirements of spot markets are feasible, and their implementation is now well developed.

The gains from trade obtained from an ISO's multilateral spot market are reduced in proportion to the extent of bilateral contracting in forward markets. Even so, gains remain because forward contracts account incompletely for contingencies and imperfect predictions of loads. Even Britain's NETA system, which relies on forward contracting up to a few hours ahead of real-time operations, conducts a real-time balancing market in which operators re-allocate supplies to follow the load, alleviate congestion, and procure supplementary reserves. Most other ISOs rely on a day-ahead market as the primary multilateral market for energy trading because it can be integrated with unit commitments and scheduling, and engineering operations can ensure physical feasibility in advance by establishing reserve assignments and alleviating transmission congestion. In the United States, FERC insists that physical feasibility is established day-ahead so that the real-time balancing market is less volatile and less vulnerable to gaming that might threaten reliability or cause extreme prices. But the choice between the extremes represented by the NETA system and the tight controls enforced in the United States evidently depends on local circumstances; e.g., the greater prevalence of transmission congestion and tighter energy supplies in some regions may account for the choice in the United States.

The efficiency of a multilateral spot market depends ultimately on whether price-mediated transactions are sufficient. They are insufficient if efficiency depends on significant public goods or other externalities. One can interpret the system operator's management of the grid as ensuring the public-good aspects of reliability. The other main externality is environmental, and in many countries it is addressed by markets for emission allowances. Even so, the more fundamental impediment to efficiency concerns the scope of wholesale markets. In principle, efficiency requires that each scarce resource has its appropriate price. To a great extent this requirement has been addressed by multiple simultaneous markets for energy, reserves, and transmission, and further, by prices for each that are differentiated by time and location; i.e. by spatially differentiated "nodal" prices established at short intervals. But a peculiarity of power markets is that there are other resources that occasionally are scarce; e.g., reactive power for local voltage support. Also, some products are imperfect versions of the resource that is actually scarce; e.g., the reserve categories reflect imperfectly the scarcity of fast-response resources, and in particular the key quality attribute, which is ramp rate or start-up time.

Schemes have been tried to establish prices for reactive power and ramp rate, but system operators usually find it sufficient to rely on engineering procedures and standard reserve categories. In the occasional instances that its markets do not provide adequate resources to manage the grid, engineers retain authority to issue dispatch directives that are settled according to rules for "out of market" transactions. A prevalent deficiency in the United States is persistent under-scheduling: when day-ahead generation schedules derived from the energy and reserve markets provide insufficient online generation to meet predicted loads, the ISO must make additional unit commitments and pay generators whatever portions of their start-up costs are not recovered from market sales. Some ISOs encourage arbitrage between day-ahead and real-time prices (by allowing "virtual" bids day-ahead that are not backed by physical resources or loads) in an attempt to reduce under-scheduling. In sum, one can conclude that from an operational viewpoint the market conducted by a system operator is inherently inferior to fully consolidated operations within a vertically integrated utility, but increasingly the prevalent view is that the disparity is small, and that such markets are mainly successful.

The sufficiency of price-mediated transactions must also be considered from the viewpoint of market participants. At a mundane level, the obvious burdens that the ISO imposes on participants (bidding, responses to dispatch directives, settlements, etc.) are likely greater than those in vertically integrated systems, and in some the ISO's expenses and therefore its grid management charges (called "uplift") are higher than anticipated. Britain's NETA system is notable for using performance-based regulation of NGC that rewards reduction of this charge. A basic deficiency of simple market clearing is that it cannot cope directly with thermal generators' nonconvex cost components such as start-up and no-load running costs, and nonconvex operating constraints such as minimum generation rates and maximum ramp rates. Since this deficiency arises partly from the definition of the traded products, alternative definitions have been proposed (e.g., enabling a base-load generator to bid to supply energy steadily over the day) but not widely adopted.

The two main alternatives are self-scheduling of generators by their owners (e.g., NordPool, NETA, California, Texas) and in those systems that inherited the operating procedures of power pools (e.g., PJM, New England, New York), central optimization of schedules for those units not committed to bilateral contracts (using "three part bids" that include the fixed-cost components and generators' reports of their operating constraints). Unit commitments by the ISO require that the portion of fixed costs not recovered from market prices is uplifted. The ISO schedules additional units to enhance reliability, but in

several systems it is alleged to depress energy and reserve prices. New England introduced separate semi-annual markets for offline fast-start reserves mainly to provide adequate revenues to attract sufficient installations of CTs needed to protect against large contingencies. As described later, a general issue is whether an ISO's markets stimulate investments in an optimal mix of generation technologies.

1.3.2.3. Forward markets

Forward markets for longer-term bilateral contracts enable both parties to hedge against price and/or quantity risks. Bilateral contracting plays a large role in all wholesale power markets due to the high volatility of spot prices. Some contracts impose physical requirements but most are essentially financial hedges against spot prices, as for example in a "contract for differences" (CFD) in which the seller and buyer insure each other against deviations of the spot price from the strike price specified in the contract. The central role of forward contracting was evident in the California crisis when the California utilities, which were largely prohibited from contracting forward, encountered severe financial difficulties while other utilities in nearby states in western United States that faced equally high spot prices were not jeopardized because they relied on spot markets for small shares of their procurements, usually less than 10%. When the state intervened to stabilize the California market its principal tactic was to secure long-term contracts that thereafter provided the bulk of the utilities' requirements.

It is important to realize, however, that long-term contracts are risky in a different way. In California the state signed contracts with generators that specified prices that turned out to be exorbitant in the long run; moreover, the contracts specified fixed quantities that in some circumstances were excessive. This debacle repeated California's earlier mistake in the 1980s when it offered QFs long-term contracts at prices that later were revealed as excessive. Although contracts written as options could have avoided these unfavorable outcomes, this experience illustrates the more basic source of the risks inherent in long-term contracts. Wholesale power markets are inherently vulnerable to systemic risks, i.e., risks that cannot be fully dissipated by mutual insurance between contracting parties. Systems with substantial hydro resources are vulnerable to prolonged droughts, those with mainly thermal plants are vulnerable to changing fuel prices and new-generation technologies, and on the demand side, both are affected by seasonal and annual weather patterns, business cycles, and other large-scale economic developments – some cyclical and some reflecting secular trends. In the regulated era, systemic risks were moderated by recovering costs from retail rates that varied slowly over extended periods. Restructuring introduced a new tension between the advantages of forward contracting in insuring against short-term spot-price volatility and the risk that the strike price or promised quantity specified in a contract would turn out *ex post facto* to be unfavorable to one or the other party.

Trading of standard intermediate-term bilateral contracts is vigorous in those systems (e.g., NordPool and Britain) with power exchanges that effectively minimize search and transaction costs. These contracts are usually for fixed quantities, which is somewhat anomalous since on general grounds one might expect other forms, such as option contracts, to be useful hedges against quantity risks. However, recent years have brought a greater variety of contract forms and tolling agreements, and some of the innovative contract forms that might be used to ensure resource adequacy.

In all systems the major share of power generation is covered by forward contracts; e.g., in PJM bilateral contracts account for about twice the volume traded in its spot market, and in Britain's NETA system they are nearly 100%. FERC's SMD supposes that the bulk

of power trading will rely on forward contracts. In particular, it aims to confine the spot market to adjustments day-ahead and real-time to address contingencies and to assure physical feasibility and reliability. The volume of real-time trading is usually under 10% in well-functioning systems – except during California's crisis when real-time trading approached 50%, presenting dire threats to reliability. Since then, forward contracting in California has been essentially mandatory, enforced with penalties for deviations from day-ahead schedules that exceed 5%, until recent changes to its market rules.

Regulators usually exempt a utility from prudency reviews for a moderate amount of purchases via intermediate-term bilateral contracts, provided they are standard contracts traded in organized markets with adequate competition and transparency. But this exemption does not apply to long-term contracts and to any transactions that hint of self-dealing with affiliated generation companies. During the initial phase of restructuring, divestiture of a utility's generation assets was eased in those systems that included so-called vesting contracts in terms of the spin-off of the generation subsidiary or sales of its assets. The assets were bundled together with contracts that fixed the prices and quantities of continued sales to the utility for several years. This procedure avoided self-dealing while providing the requisite financial hedges for the seller and buyer after divestiture. It also had a profound effect on dominant suppliers' influence on prices in spot markets.

In general, a generation firm's gain from withholding supply or bidding higher to raise spot prices is reduced in proportion to the amount of its capacity that is committed to filling the requirements of forward contracts. The market influence of dominant suppliers in the England–Wales system increased after the expiration of vesting contracts, and analogous effects are now evident in Australia. Regulatory policy has therefore often focused on measures to ensure that both generators and utilities are substantially hedged against spot-price volatility by forward contracts. A significant impediment, however, is that utilities subject to competition from other LSEs and from IPPs are reluctant to sign very long contracts in view of the risk that their service obligations might change substantially. Their role as the retail provider of last resort (POLR) exacerbates this risk; e.g., an industrial customer might opt for bilateral contracting with an IPP when market prices are low and then later opt to return to service from the utility when prices rise.

Many generation companies insist on the importance of long-term forward contracting. Their incentive to insure against spot-price volatility is greatly strengthened by effects on their costs of capital. The profitability of an investment in a new plant depends crucially on the cost of capital obtained from lenders and equity investors. Long-term contracts for major portions of the plant's capacity reduce the risk of the investment, and thereby the rate of return demanded by sources of funds; indeed, a lender often treats long-term contracts rather than the physical asset as the main security for a loan.

1.3.3. Retail service after restructuring

Cost-of-service regulation was long implemented in a way that contradicted its premise. A utility was reimbursed for costs incurred, but not until late in the regulated era was a customer charged the actual incremental cost of the service provided. Equally inefficient was the absence of differentiated service conditions that would allow customers a range of choices beyond simply the number of kilowatt-hours (kWh) to draw from the system at the standard cents-per-kWh (¢/kWh) price. Limited choice and uniform pricing were expedient in the years when the infrastructure of the electricity industry was being established, universal service was a dominant consideration, the technology of retail service

delivery was primitive, and metering and billing were major impediments to service differentiation. Nevertheless, regulators continued these policies long after the limitations that previously justified them had relaxed.

Cross-subsidization from industrial and commercial to residential customers was a basic feature of universal service, but there were others derived from technical and economic factors. The most obvious considerations stemmed from the public-good aspect of grid security and the fact that service reliability is largely uniform. A customer's service might be interrupted or curtailed, but quality attributes like frequency, waveform, and voltage are inherently uniform in a system with alternating current. Interruptions and curtailments can be imposed selectively only with costly metering and control technologies, or by direct communication that was practically confined to large industrial customers. The uniformity of quality attributes implied a basic tension between those customers who preferred lower rates for lower-quality and less reliable service (e.g., heating and cooling applications) and those who preferred higher rates for higher-quality and more reliable service (e.g., lighting and industrial production).

This tension was resolved in favor of greater reliability for several reasons. One was the technical advantage of a highly secure grid and the low cost of extending high quality to all customers, and another was the importance of high quality in promoting economic development. But the basic enabling feature was cross-subsidization that in effect charged premium rates to industrial customers for high quality that allowed lower rates for residential customers and extension of universal service. There was always an array of special provisions (e.g., low offpeak rates for street lighting and other municipal services) and special considerations (e.g., provision of the highest quality to hospitals and other essential facilities) but here we focus on the main tensions among industrial, commercial, and residential customers. These customer categories are used here as surrogates for the much more complex diversity of preferences among customers; and even for a single customer, diverse preferences in relation to different appliances and technologies (e.g., heating/cooling, lighting, production, information/communication). This heterogeneity implies efficiency gains from service differentiation, but in the electricity industry there were, and remain, severe technical and cost barriers that preclude full differentiation of service conditions and rates.

One can describe restructuring as a late stage of the more general trend to improve overall efficiency. At the retail level, this entailed unbundling of service components, pricing based on incremental cost, and, inevitably, a declining role for cross-subsidization. This trend included all the infrastructure industries, but the discussion here addresses only its effects in the electricity industry and the developments in the retail sector, and the next subsection addresses developments in the generation sector. Both cases emphasize the efficiency improvements that were sought through restructuring, and largely ignore the political resistance that inevitably accompanied the elimination of subsidies. It also ignores the role of subsidies from the government in some developing countries.[6]

[6] The major industrial countries decided long ago that the electricity industry must cover its costs from retail rates paid by customers. A major exception occurred when the state of California issued debt to fund a 10% retail rate reduction during the first years of restructuring, and then during the crisis assumed financial responsibility for the utilities' wholesale procurements, and later issued debt to fund it. National transmission companies indirectly rely on the government's good credit but costs are recovered from customers.

Economic theory predicts that a uniform flat rate causes inefficiencies for three reasons, all due to the heterogeneity of customers.

- A uniform rate does not reflect directly the incremental costs of services demanded by different customers in response to that rate; customers are more or less sensitive to the costs they impose on the system.
- A uniform rate applies to an undifferentiated commodity, whereas customers have differing preferences for the various quality attributes of service.
- A flat rate foregoes opportunities to recover infrastructure costs with less distortion of incentives.

Economists invoke two general theories about how to recover fixed costs from regulated retail rates in a way that promotes overall efficiency. One theory is due to Ramsey and its application to pricing by utilities is due to Boiteux and Mirrlees (summarized in Wilson, 1993). Generally, it supposes that the utility would run a deficit if infrastructure costs were not recovered by commodity taxes included in retail rates. Its characteristic implication is that customers with more inelastic demands should pay larger shares of the deficit. This theory was largely rejected by regulators due to the characteristic feature of retail service that demands are most inelastic among residential customers and least among industrial customers.

The other theory aims to sensitize customers to the cost implications of their demands for service. This theory has been applied in two very different ways, depending on whether costs are measured in the long run or the short run.

- A prominent application of pricing that reflects long-run costs is used in France, where tariffs are designed to emphasize the long-run implications for the utility's investments. A commercial or industrial customer pays a "demand charge" that depends on its peak load and then, for each kilowatt within that peak load, an energy charge that depends on the number of hours that kilowatt is used during the year (which requires a special meter).[7] In effect, this scheme charges the customer for its load-duration profile over the year.
- Rates that reflect short-run costs are differentiated by time or events. The simplest tariffs distinguish only between peak and offpeak periods. Equally simple in concept is real-time pricing based on the actual system marginal cost in each hour, but metering costs have confined applications to large industrial customers.

Until recently, the costs of communication or control, metering, and billing were long the impediments to differentiation of rates based on peak–offpeak periods or real-time pricing (Zarnikau, Chapter 8 in this volume). It remains one reason that rates differentiated by times or contingencies rarely extend to residential and small commercial customers, but equally important are customers' own costs of real-time communication and control, and most fundamentally, their reluctance to bear short-term price volatility and inability to obtain financial insurance.

[7] This form of retail pricing, called a Wright tariff in the United States, was used in the early years of the industry. See Wilson (1993, section 2) for an extended description of the tariffs in France, circa 1990.

1.3.4. Generation operations and investments after restructuring

Most systems allow self-scheduling of generation committed to bilateral contracts, including generation within utilities, and some allow self-scheduling for all generation. In the United States those systems that give the ISO responsibility for all other scheduling impose substantial requirements for adherence to directives for advance unit commitment and scheduling, and continuing compliance with dispatch directives. In addition, the ISO can designate plants that "must run" for reliability and pay them their incremental costs. However, these requirements differ immaterially from the previous procedures within utilities and in power pools. Even the proliferating regulatory interventions (e.g., "must-offer" obligations to bid all available capacity, and advance scheduling of deferrable maintenance) largely re-establish the comprehensive control by the utility or the power pool before restructuring – the notion that tight control is unnecessary was abandoned after the California crisis.

Following the California crisis, FERC proposed its SMD that essentially replicates the design of the northeastern ISOs descended from previous power pools. From the viewpoint of a generator, SMD really has only one product, which is generation capacity that is mostly stable from day to day and in every hour, as are its other attributes such as location and ramp rate. The SMD requires that an IPP must provide each day for each generation unit a "three-part" bid that specifies its start-up and no-load costs and its schedule of bids for energy. In addition, the ISO knows its location, maintenance schedule, heat rate, ramp rate, minimum and maximum generation levels, and other technical parameters, as well as its fuel constraints and commitments to bilateral contracts and to exports to other control areas.

If one takes FERC's SMD as the standard, then one can summarize a generator's viewpoint rather simply. Operations are about the same as they would be in a vertically integrated utility or a power pool. Indeed, because many IPPs are subsidiaries of energy companies that own both generators and utilities, the organizational aspects are unchanged in some respects. But financial matters are vastly different because remuneration derives entirely from market prices. The main effects are therefore strong incentives to minimize costs, and because they are completely exposed to the volatility of market prices, strong incentives to contract forward via bilateral contracts.

This conclusion differs somewhat for other systems (e.g., NETA, NordPool, Germany, Australia) that rely less on optimization by the system operator and more on self-scheduling, and rely more on forward contracting than on spot markets. The main conclusion remains, however, that for generators it is the financial implications of restructuring and liberalized markets that are most important. This accords with expectations that were raised when restructuring was initiated, since even then it was believed that operational aspects would not be changed materially if a system operator took over ongoing management of the grid, and possibly also spot markets. The results attest to the success in establishing well-functioning system operators (of various designs), and also to success in making profit as measured by regional market prices the criterion of financial performance.

1.4. The Unsolved Problems of Liberalized Markets

This section compiles a summary of the problems that persist after the initial years of restructuring of the electricity industry, and the accompanying liberalization of wholesale and retail markets. The discussion is divided into two subsections. The first summarizes the problems that *cannot* be solved efficiently by market processes. The second summarizes the problems that *might* in principle be solved by market processes, but that in fact current

designs have made little progress in solving adequately due to various practical aspects. In both cases these problems are deeply intertwined with the fundamental problem of how best to allocate risk bearing among market participants.

1.4.1. Problems not to be solved by markets

This subsection focuses on electricity industry functions that require continuing regulatory interventions, regardless of market design. These functions including meeting transmission system requirements, maintaining a reliable grid, and guaranteeing universal service.

1.4.1.1. Meeting transmission system requirements

The role of the transmission grid as necessary infrastructure is even more important in liberalized markets because it is the common highway for energy trading. An AC system is synchronized continuously over its entire span, and flows must be adjusted continuously to control frequency, voltage, and line loadings. Flexible AC Transmission System (FACTS) technology may eventually allow flows to be regulated by varying impedances, but most systems presently require operators to manage regional segments and to coordinate among them. The natural monopoly and economies of scale of transmission investments are therefore reinforced by the necessity of tightly coordinated operations. Recognizing this, many countries long ago established state-owned or regulated national transmission companies charged with ensuring adequate investments and ongoing operations.

Those with privately owned or predominantly utility-owned transmission systems depend instead on regulation. Most simply provide for recovery of investment and maintenance costs, while the more advanced impose grid management charges and congestion charges and eliminate pancaking of transmission access charges. A fundamental innovation of restructuring, however, is to require open access and nondiscriminatory pricing according to the principles of common carriage – minimal requirements for liberalized wholesale markets.

These measures, however, omit three fundamental requirements. One is efficient management of grid operations. This is not assured by assigning the task to an ISO of the form that, in United States, is a non-profit bureaucratic organization with diffuse incentives, unresponsive governance, and discretion restricted to explicit rules codified in its tariff. Performance-based regulation, as exemplified in the United Kingdom's regulation of the NGC, offers prospects for stronger incentives and greater discretion to manage grid operations flexibly to cope with circumstances as they arise.

The second fundamental requirement is planning of transmission expansion. Restructuring has weakened the integrated resource planning previously undertaken by utilities and in the United States left no agency responsible. This is especially serious now that liberalization has impaired coordination between transmission and generation investments. When generation investments are undertaken privately, it is imperative that transmission planning establishes reliable forecasts of the topology of the grid in future years so that decisions about generation investments can be adapted to the configuration that will be in place during at least the first 10 years of a plant's life. The best plan will likely exploit the possibilities for substitution between generation and transmission that an integrated plan might achieve incompletely, but it is still necessary to establish reliable predictions of grid expansion. Regulatory authority is often required because private generation companies invariably prefer to build in a load pocket even if transmission expansion might be more efficient.

The third fundamental requirement is financial incentive for transmission investment. The regulated rate of return usually allowed for cost recovery of transmission investments is inherently vulnerable to the "Averch–Johnson effect" of favoring excessively capital-intensive investments (Sherman, 1985).[8] This is especially true in liberalized markets because transmission projects must compete for funds with other projects, such as generation investments. A deeper problem, however, is that private incentives for transmission investments can differ substantially from social motives because the distributional effects can be substantial. For example, a retail utility in region A may want to fund a transmission line that enables it to purchase energy from region B where generation costs are lower. But if the effect of the new line is to equalize wholesale energy prices in regions A and B then an accounting of the aggregate social benefit must also consider the higher energy prices charged to retail customers in region B, as well as the effects of price equalization on suppliers in both regions. There are also new motives for transmission projects undertaken to diminish the market power of local generators.

1.4.1.2. Maintaining reliable grid operations

An important innovation of restructuring is the assignment of ongoing grid management to a system operator. The public good obtained from the transmission infrastructure depends continuously on protecting quality attributes (frequency, voltage, waveform), service reliability, and security against cascading failures. As mentioned previously, power markets are fundamentally incomplete due to disparities between technical constraints on supply (e.g., generator ramp rates, reactive power requirements) that affect reliability, and customers' perceptions that they have no choice for power of the requisite quality. Rather than relying on markets, operating procedures rely on well-developed engineering principles to sustain quality and to protect reliability and security. A system operator obtains needed resources from spot markets (or "out of market" if necessary), and it supports and facilitates these markets, but ultimately it invokes command-and-control methods to ensure rigorous coordination. This reflects the reality that grid management assures physical feasibility, whereas markets determine mainly the financial terms for settling transactions.

There remains, however, considerable latitude for the system operator to affect the efficiency of congestion management and the efficiency of its spot markets. The performance-based incentives and wide discretion allowed NGC in the United Kingdom recognize this. Indeed, the reductions in NGC's uplift and the increases in PJM's congestion cost strongly support a conclusion that incentives for the system operator should be an important ingredient of regulatory policy. In contrast, in the United States the prevalent view that engineering "best practice" determines the ISO's actions is the main justification for ignoring incentives and relying on a bureaucratic non-profit organization. FERC's orders allow ITCs, like NGC, to operate within ISOs and RTOs so one expects that in the future their role may increase, and thereby incentives might be strengthened. However, a basic impediment to the flexibility and discretion required for full efficiency remains the reliance in the United States on the rigid tariff prescribing the main aspects of each ISO's operating procedures and market rules, as well as the prohibition against stakeholder involvement in governance.

[8] Averch–Johnson effect (AJ effect), named after two economists who found that under rate-of-return regulation, if the allowed return is greater than the required return on capital, the firm will tend to over-invest in capacity beyond what is needed for economically efficient production.

Among system operators, the designs of spot markets differ markedly. Their scopes range from a single balancing market to multiple forward and real-time markets. Their procedures range from simple market clearing to elaborate optimizations of all aspects simultaneously. And their financial settlements range from energy payments to prices elaborately differentiated by time and location and differentiated among energy, various reserve categories, transmission, and in rare cases even payments for reactive power and other related products. There is no simple explanation for the huge difference between the United Kingdom's NETA, which uses only bilateral trading and self-scheduling up to a few hours before the real-time balancing market, and the enormously complex markets of the ISOs in the United States. A few features are evidently relevant (e.g., the lower incidence of transmission congestion in the United Kingdom) but both evolved from power pools with somewhat similar procedures initially.

A key difference is that, for complex reasons, the United Kingdom rejected the previous pool style of organization when it devised the simpler NETA, whereas at nearly the same time in the United States the experience of the California crisis motivated FERC to guard vigorously against a repetition by proposing its SMD. The result is that NETA relies on NGC to assure physical feasibility on a time frame of hours, while the ISOs establish physical feasibility day-ahead in consolidated markets for energy, reserves, and transmission (including unit commitments), maintain dispatch control throughout, and penalize real-time deviations. It remains unclear whether these and other system operators might converge to market designs with somewhat similar features. Continued experimentation with market designs might eventually produce convergence, but it is also possible that they will differ permanently due to initial conditions. Certainly in the United States the shock of the California crisis just 2 years after its ISO began operation has remained so vivid that comparisons with the performance of the NETA design are ignored.

The crux of the difference is the considerable confidence in the United Kingdom that NGC can manage physical feasibility and that bilateral trading suffices for efficient trading, while the accepted view in the United States is that physical feasibility must be established absolutely a full day ahead – for fear that any discrepancy will again open opportunities for gaming and abuses of market power – and therefore that the ISO's markets must be multilateral, consolidated, and optimized. Differing experiences are also relevant to comparisons among other systems – NordPool, Germany, Australia, New Zealand – each of which has a market design whose unique features were heavily influenced by experience. However, an open issue for every market is the provision of adequate investment incentives to ensure long-term resource adequacy.

A facet of these considerations is paramount when predicting the future role of system operations. One might make the case that markets in NordPool (established by the national transmission companies) and Germany (conducted by utilities) developed organically as energy trading grew, and in the United Kingdom the privately run markets for bilateral trading have developed vigorously after NETA began. But in other cases regulators and/or legislators have either chosen the market design or largely determined its main features, except for those aspects narrowly circumscribed by engineering requirements. Some may be good designs, and certainly others have been deficient – notably the deep flaws imposed on the California design by the legislature and the PUC. A basic issue to be resolved is whether in the future the spot markets of the system operator can and should cater to the commercial interests of the market participants the way other commodity markets do, or whether the peculiar technical requirements of power systems require that regulators have final authority to specify the market design.

1.4.1.3. Guaranteeing universal service

One part of retail service that is not entirely amenable to market processes is assurance of universal service. Extension of distribution lines to remote customers is now rarely a problem, so the main requirement is provision of service under standard terms and conditions. Regulators retain authority to define standard service plans, and to ensure that there is a POLR. The regulatory compact was the means in the era of vertically integrated utilities, but other means are now possible. The POLR was selected via a procurement auction in some cases where the utility's incumbency advantage was not so great as to exclude effective competition from other LSEs. But competitive procurement with fixed remuneration runs the risk that a POLR that is an unregulated LSE becomes insolvent when wholesale prices rise. As a result, the dominant mode is to rely again on the regulatory compact, so that utilities provide basic service and recover their costs from retail rates that are nearly level over time.

This mode is deeply at odds with the initial view of what liberalization of retail markets could accomplish. Ideally, retail liberalization implies that retail customers pay the hourly wholesale price for energy (in addition to other charges for distribution) but they also purchase financial hedges against price volatility to the extent they prefer. This ideal cannot be realized anytime soon because metering is still primitive and the markets for financial hedges are undeveloped, but more basically this ideal is impractical for the (mostly small residential) customers most affected by universal service, even if financial hedges were subsidized. Relying instead on utilities, and using the regulatory compact to assure cost recovery over time, is likely to remain the only practical solution for decades.

This poses two basic problems. The first is that the "core" of customers opting for basic service is inherently unstable. Those who opt out when wholesale prices are low are equally motivated to opt again for basic service when wholesale prices are high. Charges might be imposed on those who leave the core and again on those who return, but exactly how such a system would ensure financial stability through prolonged swings of wholesale prices is not well understood and has not been fully developed.

The second basic problem occurs even if those who opt out of the core never return. The customers who are least costly to serve (e.g., those with flat load profiles) are precisely those who will be offered the most attractive terms by IPPs and non-utility LSEs, and therefore they are the ones most likely to opt out of the core. This leaves in the core only the customers most expensive to serve, implying that their retail rates (even if leveled over time) will rise as the core is depleted of the more profitable customers. This scenario is entirely realistic since it merely repeats the dire experience of those utilities that, late in the regulated era, were subject to bypass by profitable industrial customers who opted for self-generation or co-generation, or later, direct contracting with IPPs.

1.4.2. Problems not yet solved by markets

The discussion now turns to a different class of problems, those that liberalized markets were initially believed to solve, but in fact have not. These include assuring adequate generation resources and benefiting retail customers, an issue also addressed in Chapters 9–13 of this volume.

1.4.2.1. Assuring adequate generation resources

The generation sector was thought to be the most amenable to market solutions. There is some evidence that operating efficiency has improved at divested plants, as measured by labor and fuel (heat rate) inputs. Other aspects of operations are less clear in those systems

that exclude self-scheduling for those plants not committed to bilateral contracts. Those ISOs in the United States that use variants of FERC's SMD are most extreme in abandoning market-clearing processes in favor of comprehensive optimization, including unit commitments and scheduling of energy generation and reserve assignments. Although these ISOs settle accounts using prices computed as the marginal costs of unbundled products, their operations consist essentially of centralized allocations of available generation and transmission capacities. These procedures differ little from those used previously in vertically integrated utilities and power pools. These ISOs in the United States might eventually resemble the more decentralized systems in other countries if, as FERC intends, bilateral contracting becomes more pervasive. At present, however, the fraction of plants scheduled centrally is substantial and stable, and in many ways FERC's insistence that each ISO must ensure day-ahead physical feasibility necessitates centralized unit commitment and scheduling, and continuing dispatch control until real-time.

The mix of generation investments has been problematic in most ISOs. An encouraging sign is that many systems now have substantial dispatchable loads that compete with generators in reserve markets, and increasingly the demands of industrial customers respond to real-time prices. The number of interruptible or curtailable retail service contracts remains comparable to the pre-restructuring period, but in general the utilities' active promotion of demand-side management declined after restructuring as regulators withdrew the subsidies and incentives previously provided. Regulators have continued some subsidies of generation from renewable sources, often by using market mechanisms such as auctions, but also by innovative schemes such as the tradable "green certificates" introduced in Europe. In addition, renewable sources increasingly compete on equal terms, albeit with inherent disadvantages because generation need not be dispatchable and can be intermittent (e.g., wind and solar).

Now the mix of generation investments is also affected by requirements for tradable permits for emission of pollutants, such as sulfur dioxide and nitrous oxides. The ratification of the Kyoto Treaty will in future years lead to comparable requirements for emissions of carbon dioxide and other greenhouse gases.

The most severe problem concerns investments in peakers (e.g., CTs) that routinely provide fast-response offline reserves and occasionally generate to meet peak loads when prices exceed their high marginal costs. In other countries such as Australia, bid or price caps are sufficiently high that peakers obtain substantial revenues in times of supply scarcity (and encourage demanders to obtain financial hedges). While a high price cap is arguably the best solution, the low caps in the United States curtail CTs' revenues from occasional generation and force them to rely mainly on payments for reserve assignments. However, centralized unit commitment by the ISO, usually undertaken to strengthen reliability, has the side effect that it depresses the price for offline reserve capacity, because the minimum operating level of a newly committed unit tends to create an excess supply of spinning reserve.

The usual motive for extra unit commitments is under-scheduling of loads in the day-ahead market, compared to the ISO's forecast of the next day's peak load. This problem is acute in New England because its ISO needs substantial offline reserve capacity to meet its unusually large contingencies, but reserve prices are insufficient to cover the carrying cost of a CT. The ISO there is exceptional for addressing this problem by allowing CTs to capture scarcity rents. One device is a separate semi-annual market for offline reserves, and a second is a contingent capacity payment awarded to all available capacity whenever reserves are in short supply in real-time operations.

These measures are indicative of a general trend toward special provisions intended to obtain sufficient supplies of reserves in daily operations and more elaborate efforts to assure adequate generation resources. The plain fact seems to be that adequate reserve capacity – both fast-response reserves in the short term and total available capacity in the long-term – is not assuredly provided by the financial incentives derived from wholesale markets. Reserve capacity can be insufficient when prices in these markets are depressed by regulatory interventions, as in the United States where low price caps and residual unit commitments for reliability affect revenues obtained by CTs. But the basic economic problem is that individual customers have insufficient incentives to pay for reserve capacity (and its chief attribute, a high ramp rate) that is idle most of the time and that serves mainly to provide the public goods of system reliability and grid security. Customers and regulators are acutely sensitive to these matters when prices are high because loads are high or supply is scarce, but ordinarily their financial incentives are myopic compared to the integrated resource planning previously conducted by utilities.

A chief motive for separating the organizational components of vertically integrated utilities was to establish stronger incentives for efficient investments in generation. Rather than assured cost recovery obtained by a utility, a private investor in generation obtains the subsequent rewards and bears the risks that follow from its decision. But the evidence is mixed as regards investment. Success stories in countries like Australia and in ERCOT, which is not subject to FERC jurisdiction, are matched by a bleaker picture in other parts of the United States and in Europe. The insufficiency of new investments was initially attributed to prolonged regulatory uncertainty, compounded by the demise of new contracts for QFs, including both co-generation and those using renewable energy sources. Then high wholesale prices in the period surrounding the California crisis stimulated substantial investments by IPPs, mainly in combined-cycle gas-fired plants. But this was followed by rather low prices that resulted in financial distress for many IPPs and bankruptcy for some. This could be a sign that boom and bust cycles are endemic in the generation sector, as they are in some other commodity industries. But if so then it reflects a basic deficiency in implementing liberalized markets, which we now address.

Restructuring was based on two premises implied in the seminal book by Joskow and Schmalensee (1983). One was that the vertical integration of utilities could be replaced by bilateral contracts between generators and large customers, or with retail utilities and other LSEs, assisted by multilateral markets for spot trading. The other was that generators could obtain capital on comparable terms directly from financial markets without relying on inclusion within a regulated utility with its assured cost recovery. Even California, which prohibited long-term bilateral contracting by utilities during the 4 years allowed for recovery of stranded costs, presumed that contracting would ultimately prevail. Contracting was certainly the dominant mode initially in those systems that used vesting contracts to smooth the first years of transition. In some systems (e.g., NETA, NordPool, Australia) both long- and mid-term contracting are vigorous, and others (e.g. France, Alberta) stimulated contracting by auctioning power procurement agreements (PPAs) sold by incumbent owners of large amounts of generation who were not required to divest ownership and management of their plants.

But experience has shown that there are basic problems involved in extending the role of contracting. As described earlier, sellers and buyers have a natural incentive to insure others against volatile spot prices, since these are just financial transfers between them. The prevalence of mid-term CFDs reflects this mutual incentive for price insurance. Other stipulations involve the contract duration and the quantities involved. Utilities and other LSEs are hesitant to contract long-term for fixed quantities when their customers can switch

to alternative providers, while generators prefer such contracts. Similarly, utilities and other LSEs want to adjust contracted quantities to contingencies affecting loads whereas generators do not. These considerations may account for the smaller role of contracting in the United States.

The creditworthiness of the contracting parties also seems to have been an impediment in the United States, both in the California crisis when the utilities were financially distressed and later when there were several bankruptcies of major generation companies. Utilities' regulators are wary of default on supply contracts, and they are also reluctant to approve utilities' purchases of purely financial short-term contracts (e.g., futures contracts, as opposed to forward contracts that entail delivery). Indeed, most PUCs prohibit their inclusion in cost recovery on the grounds that they amount to speculation or arbitrage, a view that is evident also in federal legislation that requires separate and more intrusive regulation of exchanges for commodity futures.

Also relevant is the slow development of innovative forms of forward contracts. For instance, an option contract can enable an LSE to "call" for the delivery of power supply at a pre-specified strike price in contingencies when the spot price exceeds the strike price. An option contract enables a buyer to hedge against high prices without exposure to the "volumetric" risk that the contracted quantity exceeds the amount required to serve its load. A multitude of other contract forms are also possible; e.g., a "spark-spread" contract enables a seller to hedge differences between fuel and power prices. Predictions that contracting would develop vigorously, and fund investments in generation, were predicated on the assumption that adequate contract forms could be developed to overcome the concerns of sellers and buyers, and that prudency reviews and creditworthiness would not be major problems.

Predictions that contracting would replace vertical integration have been only partly realized because in fact the impediments to contracting interacted with the other major premise of restructuring – namely, that generation companies could obtain capital directly from loans and from financial markets for bonds and equity shares, much like suppliers in other deregulated industries (transport, telecommunications, gas). In the United States the chief models were gas transmission companies, which typically obtained funds at low cost because before construction of a pipeline most of its capacity was pre-sold as firm transmission rights under long-term contracts, often for durations as long as 20 years. Some IPPs have pre-sold capacity for durations long enough to provide security for loans to fund investments, but the overall pattern is mixed. In California, over 90% of the 8 GW of new capacity (an increase of 20%) installed in the years after the crisis was being financed by the long-term fixed-price, fixed-quantity contracts that the state purchased when it intervened. The state later found itself burdened with contracted prices and quantities that turned out to be excessive. This salutary lesson now encourages PUCs to restrict the durations of utilities' contracts for resource adequacy, and further, to impose stringent conditions for taking over the plants of a supplier in default. Default risks are substantial because of the inherent volatility of wholesale prices, even over extended periods. Recent years of sustained low prices have jeopardized the financial viability of even the major generation companies, especially those active in energy trading with exposed positions (e.g., Calpine, Dynergy, Enron, Mirant, Williams).

In principle, a generation company could be fully hedged by long-term fixed-price, fixed-quantity contracts and therefore be immune to long swings of wholesale prices, and if its counterparty is a utility with assured cost recovery, then its credit is essentially comparable to the utility's credit. But complete hedging is rarely available, and the contract duration is typically no more than 3 years, a minor fraction of the life of the plant. Financial

hedges (e.g., futures contracts) are intrinsically short-term because no counterparty to a long-term financial contract is creditworthy unless it has physical resources as a collateral. A forward contract for physical delivery over a long duration can be viable for the seller, but no LSE other than a utility is a creditworthy buyer. As mentioned earlier, the LSEs, utilities, and industrial customers must guard against quantity risks. An option contract mitigates the buyer's quantity risk, but equally it imposes on the seller the quantity risk that over long periods with low prices the option will not be called. Because the option will be called only when spot market prices are higher than the strike price, the seller serves mainly as insurer of the buyer with no comparable insurance for itself.

The result of these difficulties is that two basic premises of restructuring have been fulfilled only partly. Contracting is not as pervasive as expected, and contracting obtains only part of the financial advantages of vertical integration. Most important is that IPPs are risky companies and therefore they have higher costs of capital – and during the recent period of sustained low prices, some of the most prominent are financially distressed, bankrupt, or reorganized after bankruptcy. The ultimate consequence is that investments in new generation by IPPs are substantially impaired. For example, Calpine obtained regulatory approvals for siting and construction of three new plants in California for which it did not obtain investment funds.

1.4.2.2. Benefiting retail customers

Restructuring and liberalization of retail markets were expected to benefit customers substantially. For example, California's announcement of its decision extolled at length the merits of differentiated services that customers would obtain at competitive prices from LSEs and from utilities freed from the confines of standard service plans. Much of the gain was supposed to come from service plans that rewarded customers' efforts at demand-side management of their energy consumption, complemented by integration (called "convergence") of energy services with other basic services – especially telecommunication and (remote or local) automated control of appliances, which have yet to materialize to any great extent.[9]

Large industrial and commercial customers certainly obtained advantages from new freedom to contract bilaterally with IPPs and to purchase directly from retail energy providers (REPs) or wholesale markets.[10] But LSEs made slight inroads into commercial and residential retail markets in the United States, and subsequently, those who survived financial stresses during and after the California crisis withdrew from these markets.[11] The utilities retained 90% or more of the residential market in most states simply because few customers considered switching energy suppliers. An exception is ERCOT where retail competition is thriving and 50% of the load has switched away from their local utility suppliers. A significant fraction of commercial customers chose service from LSE's whose rates were gauged to each customer's load profile, but this was perhaps the only significant differentiation and succeeded mainly because it provided lower prices to customers less

[9] California's 1994 restructuring decision devoted two pages to a section entitled "The Convergence of Telecommunications with Electric Service Promotes Direct Access" (pp. 21–23).

[10] For example the University of California at Berkeley signed a supply contract with ENRON that guaranteed a minimum of 5% saving over the PG&E regulated rate. During the energy crisis ENRON attempted to renege on that contract and force the university to return to PG&E, but lost its bid in court.

[11] California suspended "direct access" at the end of the crisis, and the suspension continues still.

costly to serve. More elaborate differentiation was impeded by the prevalent absence of interval meters, and the costs of more elaborate meters and billing procedures.

Prior beliefs that customers would actively monitor and control usage to reduce their costs under new service plans were falsified by the same practical obstacles that had limited the results obtained previously from utilities' demand-side management programs. Equally falsified was the expectation that a majority of customers would willingly bear short-term price volatility (or could and would purchase financial hedges) in order to obtain lower prices on average. In fact, when wholesale prices were passed through to customers (in San Diego in 2000 as the California crisis began), the howls of outrage from customers prompted immediate intervention by regulators who canceled provisions allowing the utility to pass through its wholesale procurement costs.[12] A particular example of misguided expectations was the prediction, obtained from surveys of mainly residential customers, that 30% would willingly pay a premium of 10% for "green" power from renewable sources. In fact, those subscribing to green power never exceeded 3%.

More successful liberalizations of retail markets do not rely on visions of elaborately differentiated services. Instead, they concentrate on competitive retail markets as means of lowering prices for standard service plans, and for enabling negotiated rates based on a customer's load profile. Alberta's auction of PPAs, for instance, enables LSEs who purchase them to compete against the incumbent utility. Texas encouraged the entry of competitive REPs by requiring utilities to refrain from price competition with the REPs at prices below a "price to beat" until 2007 or until the utilities' market share dropped to 40%, whichever occurred first.

A basic lesson from recent years of liberalized retail markets is that only the market for large industrial and commercial customers is sufficiently developed presently to benefit fully from liberalization (see Joskow, 2005). Extending retail competition to other market segments requires major advances in infrastructure (especially metering), and fundamental redesign of differentiated services. Also required are measures that allow LSEs to enter successfully, to establish significant market shares, and to remain financially viable through periods of high prices. The reason is, in part, that they insure their customers against wholesale price volatility, although on the timeframe of a month or a year, which is shorter than utilities provide. Most fundamental is the lesson that most customers in the commercial and residential segments are deeply averse to price volatility, and equally reluctant to undertake the measures required to continually monitor and control usage. These customers will remain in the "core" served by the utility using fairly standard service plans (only slightly differentiated, say, by peak and offpeak periods) with rates leveled over extended periods. Only if there are major advances in providing them with financial hedges against price volatility, and with compensation for service interruption or curtailment, are these customers likely to depart from the core.

[12] Similarly, Ontario canceled liberalization of retail markets after prices rose 30% in the first months. Customers' attitudes about pass-through of wholesale electricity prices have no easy explanation. During the California crisis there was widespread anxiety about their monthly bills, but in fact throughout the crisis electricity rates were regulated and rigidly fixed. The anxiety can be explained in part by the fact that gas prices were passed through to customers and bills from PG&E and SDG&E included charges for both electricity and gas. But why many customers did not notice that it was higher gas prices that accounted for their higher monthly bills is mysterious. The new role of electricity as necessary for commerce and for personal well being – even TV – may be as good an explanation as any for these attitudes.

1.5. The Allocation of Risk Bearing in Liberalized Markets

This section analyzes risk management in the electricity industry and indicates the continuing role of regulatory policies and regulated retail utilities in mitigating risks for a segment of retail customers. Rather than the simplistic version of liberalization that aims simply to break apart the vertically integrated utilities, and to unbundle the products traded in markets, there is a continuing role for utilities in providing inter-temporal smoothing of retail rates, and in lowering the cost of capital by reducing their financial exposure.[13] This role requires redesign of utility investments and contracting within the context of liberalized wholesale markets, and redesign of retail rates within competitive retail markets. First reviewed are the risks peculiar to the industry, and then the institutional arrangements that have been or can be used to obtain an efficient allocation of risk bearing.

1.5.1. The basic risks affecting the electricity industry

Much of the theory and practice of risk management is based on diversification. Insurance companies are financially viable because they aggregate many small independent risks to life, health, property, etc. They can operate with relatively small reserves of equity capital because their aggregate risk is small on a per capita basis, that is, when divided relatively equally among all participants. Another form of diversification is seen in a mutual fund. Its earnings are less volatile than the earnings of the companies whose shares it owns because its portfolio is divided among the shares of many companies. Again, the aggregate is less risky per dollar invested because the companies' earnings are imperfectly correlated.

Diversification is also very important in electricity markets. For instance, the aggregate load on a per customer basis is more stable than the loads of the customers individually because the aggregate is composed of individual loads that are imperfectly correlated. Thus, for each hour of the next day, a system operator's day-ahead prediction of the average load per customer is more accurate than its predictions of the loads of individual customers. Similar considerations apply to other contexts; e.g., the system operator uses the transmission system to compensate for equipment failures in one area by drawing on resources in other areas. Analogously, the system operator can meet a high load in one area by drawing on resources in other areas that are not affected by the same weather conditions. Energy trades also take advantage of differences between seasonal variations among regions; e.g., California buys power from the Northwest to meet summer daytime air conditioning loads, and sells power to the Northwest to meet winter nighttime heating loads. The mix of generation technologies that is most cost-effective in meeting a particular load-duration profile is really the solution to a risk management problem in which diversification of generation investments is optimized.

Both spatial and temporal diversification, however, are inherently limited in the electricity industry. Investments in generation and transmission facilities require years to complete. After construction, they are irreversible, specialized, immobile, and long lived.

[13] One problem underlying the spotty record of power industry restructuring so far is that the economic theory used to justify functional unbundling of utilities has not proved as useful as originally expected. This theory, which emphasized the importance of transaction costs, depended largely on assumptions that have become outdated because of innovations in technology and market design. It now appears that many of the economic benefits sought through unbundling can better be attained through wider use of risk management contracts, obviating a compelling reason why restructuring should begin with the irreversible task of vertical unbundling of the supply chain.

In contrast, loads vary greatly on much shorter time frames. Some variation in loads is predictable and cyclical, such as the typical variations over the hours of a day, and over the seasons of the year, so the mix of generation technologies can be designed to minimize total costs over the cycle. On the supply side too there are regular patterns of downtime for maintenance, and to a substantial extent the average frequency of equipment outages is predictable. But the electricity industry is also affected by large and relatively unpredictable variations that occur over wide regions and/or over long timescales. These bring risks that cannot be mitigated by various strategies of diversification based on averaging over people, locations, or time in a cycle. These are called non-diversifiable or "systemic" risks.

Among systemic risks, the most extreme event is collapse of the grid due to cascading failures. Systems that rely heavily on hydroelectric sources are vulnerable to prolonged droughts that curtail water storage behind dams, and those that rely on fossil fuels are vulnerable to eras of high prices. Over long periods, load patterns trend away from the aggregate load-duration profile and the spatial distribution used initially to justify generation and transmission investments. Technical change can also render a generation plant or transmission line inefficient compared to subsequent investments in newer technologies.

For a casualty insurer, the analog of a systemic risk is the rare storm (e.g., hurricane) or an earthquake that devastates an area and requires that compensation be paid simultaneously to many victims of the same event. In other words, the injuries to insured customers are perfectly correlated due to their common dependence on the single event of the storm. Because a single event can exhaust its financial reserves, a casualty insurer often excludes coverage of systemic risks. For instance, a farmer can purchase insurance against damage to crops by hail (which occurs locally and briefly) but cannot purchase insurance against drought (which is widespread and prolonged). Casualty insurers typically diversify further by re-insuring a portion of their risks with other companies that specialize in aggregating calamitous risks on a global scale. Such strategies might conceivably be invoked in the electricity industry, but presently the scale of the financial risks of major events in the electricity industry is so large, and can extend over such long times that it is often the state that ultimately bears the cost – as when the state of California intervened to purchase power for utilities during the crisis.

Systemic risks are addressed in many different ways. Provision of reserve generation capacity is the first defense: it is costly to build and maintain generators that are idle most of the time, but over the long run their ultimate value is realized in the occasional events when they are called to produce power. As with an insurer, a second defense is a reserve of equity capital that can be drawn down to pay extraordinary expenses. Again, it is costly to retain financial reserves that, in effect, are idle until used to meet unexpected expenses. Like an insurer, a utility can use the third defense of re-insuring its financial risks in various ways, such as long-term contracts that transfer some portion of its risk of high wholesale prices to its suppliers, and retail service plans that transfer some risk to customers. Risk associated with moderate weather fluctuations can also be diversified through instruments such as weather derivatives that enable risk sharing among industries that are affected by weather in complementary ways. But some entities must ultimately bear the costs of rare extreme events, and in the electricity industry the prospects that financial contracting can entirely diversify systemic risks are limited.

The problem stems partly from the attributes of physical assets, which are built slowly, and are irreversible, specialized, immobile, and long lived, and, thus, largely inflexible in dealing with contingencies that occur on large scales of space and time. A contract that transfers risk from a regulated utility to an unregulated generation company leaves the company exposed to the risk but with limited flexibility, since it cannot redeploy or

quickly expand its assets to cope with events as they develop. Ideally, a long-term fixed-price contract protects the utility against sustained high spot prices (while forgoing high profits for the company), and protects the company against sustained low prices (while foregoing lower procurement costs for the utility).

The advantages for the company are inherently greater since its investments require irreversible commitments measured in decades, while the utility's advantages are confined to the durations of extreme events. But this mutual insurance against price variations is not sustainable over prolonged periods that jeopardize the financial viability of either party – as evidenced by the bankruptcies of generators in the United Kingdom and the United States during the recent period of low prices, and before that the financial distress of utilities and other LSEs during a period of high prices.

The problem also stems partly from high correlation between prices and quantities; i.e., prices are high when loads are high. A contract that provides price insurance is mainly financial, since it fixes the terms of trade, and brings mutual advantages to buyer and seller by eliminating price volatility. But a contract fixes the quantity to the benefit of the seller only by removing the buyer's flexibility in procuring the amounts needed to meet its load in each event. If the amount is too large or small, then the surplus or deficit must be corrected by spot sales or purchases. Basically, the seller wants to insure its flow of net revenues (especially if it is burdened by debt) and the utility wants to insure its total cost of procuring supplies to meet its varying load. Some contracts address these considerations directly, notably tolling contracts (sometimes called "virtual capacity" contracts) and some PPAs, in which essentially the seller is remunerated continuously for its capacity availability and operation. Meanwhile, the buyer dispatches the plant as needed and pays variable generation costs, chiefly for fuel. This achieves the primary goal of restructuring, which is to make generation companies bear the consequences of their investment decisions and to strengthen their incentives for efficient operating practices, while also enabling the utility to obtain scheduling flexibility.

Restructuring assumed that the strategies described above – physical reserves, financial reserves, and contracting – would suffice in the new era of liberalized markets. Each strategy has limitations and costs that have became clearer as experience accumulated. Remarkably, restructuring has often abandoned the traditional means of risk management, namely cost-of-service regulation of utilities. Risk was borne ultimately by retail customers, but only by amortizing recovery of a utility's accumulated costs over time so that its retail rates were substantially level. This strategy uses diversification in two ways: it distributes risk bearing widely among customers and over time. It also removes most risk bearing by the utility and its suppliers so that their capital costs are low, and eliminates the creditworthiness problem. It depends implicitly on the good faith and credit of the state in fulfilling the regulatory compact; indeed, it may be that the state is the only entity that can provide credible assurance that costs will be recovered later as promised.

This strategy also has limitations and costs. These became very clear at the time of restructuring, since regulators had often seen evidence that utilities' incentives for cost minimization were weak, and investments were rewarded on the basis of "tonnage of money invested."[14] California's initial consideration of how to restructure proposed a scheme in which utilities would continue to serve core customers much as they had

[14] This quote is from the February 1993 staff report to the California Public Utilities Commission by its Division of Strategic Planning, "California's Electric Services Industry: Perspectives on the Past, Strategies for the Future", page 100.

previously, while allowing non-core customers to purchase their power supplies directly.[15] Although California rejected this scheme, other states retained a central role for utilities and continued the practice of amortizing costs over time to level retail rates. This necessarily implies that core customers ultimately pay the full cost of the services obtained, and in particular they collectively bear the systemic risk that the cost of service for the core will be high due to extreme events and long-term trends.

1.5.2. Institutions for risk bearing in the electricity industry

Restructuring was a response to many considerations. One was the reduced role of economies of scale in generation as smaller combined cycle units became cost-effective. This and other technical advances obviated the dominant role of retail utilities on the supply side and opened prospects of bilateral contracting between IPPs and large industrial and commercial customers. For those systems not already organized as power pools, there were potential operational gains from regional operations and trading based on open access to transmission. Naïve expectations that service differentiation would proliferate amid vigorous retail competition were realized only partly, and mainly for large industrial and commercial customers. Some PUCs emphasized reduced costs of contentious procedures and litigation required to implement cost-of-service regulation, but in fact this outcome was precluded by the continuing dominant role of retail utilities in serving core customers.

A primary goal of restructuring was to strengthen incentives for efficient operational and investment decisions. Cost-of-service regulation is inherently a kind of insurance for utilities, since it guarantees to a utility that its costs accepted as prudent and accepted into its rate base are eventually recovered in full from retail rates on an amortized basis that includes the cost of capital. Insuring utilities' cost recovery was very effective in reducing the cost of capital, since their bonds and shares carried negligible risks of default and provided steady payments of interest and dividends. But inevitably, insurance dilutes incentives, since a utility does not bear the costs that result from its investment decisions and operating practices.

Restructuring apparently succeeded as regards cost reduction in daily operations, but it has had mixed success in investments, and the unfavorable consequences for the riskiness of utility shares was not anticipated.[16] Restructuring envisioned that supply-side contracts with independent suppliers, supplemented by spot market purchases, would supplant in-house generation by vertically integrated utilities. And on the demand side, differentiated service plans would include risk bearing by customers (possibly hedged by financial instruments) on a more nearly current basis than the previous regime of rates leveled over extended periods. Also expected was that customers' loads would become more sensitive to market prices. These two developments on the supply side and the demand side were

[15] The reasons for rejecting this alternative are not entirely clear. One explanation is that California rates had already reached averages 30% to 50% above the national average while relying on cost-based regulation, and they might rise further if the most profitable customers opted out of the core. But the PUC's decision focused also on predictions that complete liberalization would bring benefits from service differentiation among competing LSEs that in fact did not materialize. In 2004 the PUC reinstated the policy of promoting utilities as the provider of services for core customers. This reversed the decision made a decade earlier.

[16] Regulators did not expect utilities to be in financial jeopardy, but equity markets reacted differently. After the California PUC's decision in 1994 prices of the utilities' shares declined about 25% over the next few months.

expected to leave retail utilities with moderate commercial and financial risks, much like other firms in commodity industries – analogies to successful deregulation of the transport, telecommunications, and gas industries were often cited.

The limits to the success of restructuring regarding investments include effects of imperfect planning, especially as regards the mix of technologies and the provision of adequate reserves. But the main limit is due to imperfect contracting that leaves generation companies exposed to substantial risk and therefore required to pay higher costs for capital. The failure of demand-side innovations to develop is now viewed as fundamental.

This perception may change gradually through a long process of developing retail markets – expansion of sophisticated metering and redesign of marketing strategies – but the basic impediment is the absence of adequate financial instruments for small customers to hedge against price volatility. A utility retains its obligation for universal service as the provider-of-last-resort, and its standard service plans include leveled rates that effectively insure against short-term price volatility. Most small customers therefore choose to remain in the core served by the utility. (In the United States, those who opted for service from alternative LSEs were sent back to their utilities for default service when wholesale prices rose in the period 2000–02.) There is no evident substitute now for the retail financial services of utilities, and there may indeed be no credible substitute for the state's guarantees of universal service and of cost recovery via leveled rates. In retrospect, the anticipated gains from service differentiation for small customers must presently be seen as secondary compared to the gains from leveled rates for those in the core.

From an economic viewpoint, there are persuasive arguments that the most efficient allocation of risk bearing in the electricity industry has core customers paying leveled rates for amortized cost recovery. Except for some economically disadvantaged customers, they should pay the full cost of service in the long run because costs vary with usage. Since electricity is used universally, they should pay directly for service rather than rely on distortionary taxation by the state to cover deficits. Some industrial and commercial customers can bear short-term price volatility without difficulty, and therefore they can pay spot prices and/or contract directly with suppliers. But if other customers are deeply averse to short-run volatility then ideally they should pay level rates that recover their costs over time.

Inter-temporal smoothing of rates might be achieved by well-developed markets for financial instruments for hedging against price variations and for insurance against the consequences of curtailed service. Alternatively, cost-of-service regulation provides smoothing of rates. The choice between these two approaches depends on how seriously systemic risk limits the market for financial instruments, and how serious are the deficiencies of cost-of-service regulation in providing strong incentives. The defects of cost-of-service regulation were well known before restructuring, but the slow and ultimately inadequate development of competitive markets for financial instruments was not expected. It was also thought to be a secondary consideration, since customers retained the option to rely on core service, and in any case little or no attention was given to the problem of systemic risk. For instance, documents and orders of the California PUC and state legislation mention only short-term price variation, with no recognition that extreme events like those that initiated the later crisis might affect the restructured industry.

Most firms in the financial services industry were well aware of the threat posed by systemic risks. They hesitated about offering long-term financial instruments because they had no physical hedges. A firm that offers financial hedges against high wholesale prices runs the risk of ruin unless it can compensate by simultaneously profiting from selling power at high prices. Several firms within the energy industry became active traders and

arbitrageurs (including Dynergy, Enron, Mirant, and Williams) and they claimed that their portfolios of long and short positions were well hedged against extraordinary events. But in fact, among the major trading companies in the United States, the only ones solvent after the California crisis and ensuing events at the national level were those that had the foresight to liquidate their positions and close their trading operations in the early months of the calamity.

A summary view of these deficiencies on the demand and supply sides is that both are instances of insufficient development of auxiliary markets for financial instruments and contracts that hedge against risks. Diversifiable risks are allocated inefficiently when financial markets are poorly developed, and more seriously, non-diversifiable risks can jeopardize the entire industry as financial distress affects many participants. Dire scenarios were not envisioned when restructuring began, but they became worrisome concerns after episodes in several countries, and then became crystal clear during the California crisis, and later in some other countries such as New Zealand.

The view now is that even after restructuring and liberalization there remains a valuable role for retail utilities that use financial reserves from capital markets to smooth cost recovery over time for those customers who opt to remain in the core. The importance of strengthening incentives is also recognized, and therefore new regulatory policies are required. Remuneration of utilities via simple cost-of-service regulation must be replaced by a scheme that enables a utility to insure core customers against short-term price volatility, while also rewarding the utility for efficient operations. Chao et al. (2006) address these matters in detail, including the role of performance-based regulation of utilities.[17]

1.6. Conclusions

The argument for vertical integration in the electricity industry and also the argument for restructuring based on unbundling of its products and organizations in favor of market mechanisms are both deficient. The notion that all is needed is unbundling of the electricity supply chain and establishment of efficient short-term trading institutions, while long-term contracting and markets for financial risk management instruments will emerge spontaneously, was naïve. In retrospect, cost-of-service regulation and vertical integration of generation and retail service continues to be a powerful means of risk diversification. The extremes of vertical integration and liberalized markets are inferior to a balanced mixture of the two approaches. While unbundling may benefit large industrial and commercial customers that are able to absorb the inherent risks in the electricity supply chain, efficient management of these risks requires that restructuring retains universal service for the core of non-industrial customers who rely on regulated rates smoothed over time to recover the costs of service.

Acknowledgment

The chapter is based on research sponsored by Electric Power Research Institute. The opinions expressed in this chapter are those of the authors and do not represent the positions of any of the organizations with whom the authors are affiliated. Any errors and opinions are solely the responsibility of the authors.

[17] In its 1994 decision the California PUC explicitly provided for performance-based regulation of utilities that continued to serve core customers, but this was not implemented due to the restrictions imposed to enable utilities to recovery the "stranded" costs of previous investments.

References

Barker, J. Jr., Tenenbaum, B., and Wolf, F. (1997). Governance and regulation of power pools and system operators: An international comparison. World Bank Technical Paper No. 382.

Blumstein, C., Friedman, L., and Green R. (2002). The history of electricity restructuring in California. CSEM Working Paper 103, Berkeley, CA: University of California Energy Institute.

Bonbright, J.C. (1961). *Principles of public utility rates*. Public Utilities Reports, Incorporated.

Borenstein, S. (2004). The long-run effects of real-time electricity pricing. CSEM WP-133, Berkeley, CA: University of California.

Borenstein, S., Bushnell, J., and Wolak, F. (2002). Measuring market inefficiencies in California's restructured wholesale electricity market. CSEM Working Paper 102. Berkeley, CA: University of California Energy Institute.

California Public Utilities Commission (1993). *California's electric services industry: perspectives on the past, strategies for the future*. San Francisco, CA.

Chandler, A.D. (1969). *Strategy and Structure: Chapters in the History of the Industrial Enterprise*. The MIT Press.

Chao, H-P, Oren, S.S., and Wilson, R.B. (2006). Alternative pathways to electricity market reform: risk management approach. *Proceedings of the 39th Hawaii International Conference on Systems Sciences* HICSS39. Kauai, Hawaii, 4–7 January.

Chen, M., Cho, I-K., and Meyn, S. (2004). Reliability by design in distributed power transmission networks. University of Illinois, Urbana Champaign.

Correljé, A.F. and. De Vries, L.J (this book). Hybrid electricity markets: The problem of explaining different patterns of restructuring. Chapter 2.

Devine, W.D. Jr. (1983). From shafts to wire: Historical perspective on electrification. *J. of Econ. Hist.*, **43**(2), 347–72.

EPRI. (2002). *Review of the Current Status of Power Market Reforms in the United States and Europe*. Palo Alto, CA.

Federal Energy Regulatory Commission (2002). *Standard Market Design Proposed Rulemaking*. Washington, DC.

Hunt, S. (2002). *Making Competition Work in Electricity*. Wiley.

Joskow, P. (1997). Restructuring, competition and regulatory reform in the U.S. electricity sector. *J. of Econ. Pers.*, **11**(3), 119–38.

Joskow, P. and Schmalensee, R. (1983). *Markets for Power: An Analysis of Electrical Utility Deregulation*. MIT Press.

Michaels, R. (2006). Vertical integration and the restructuring of the U.S. electricity industry. *Pol. Analy.*, No. 572, 1–32, published by the Cato Institute, 13 July.

Newbery, D. (2006). Electricity liberalization in Britain and the evolution of market design. In *Electricity Market Reform: An International Perspective* (F.P. Shioshansi and W. Pfaffenberger, eds). Elsevier.

O'Neill, R., Helman, U., Hobbs, B., and Baldick, R. (2006). Independent system operators in the USA: History, lessons learned, and prospects. In *Electricity Market Reform: An International Perspective* (F.P. Shioshansi and W. Pfaffenberger, eds). Elsevier.

Rudnick, H., Barroso, L.A, Skerk, C., and Blanco, A. (2005). South American reform lessons – twenty years of restructuring and reform in Argentina, Brazil, and Chile., *Pow. and Energ. Mag., IEEE*, **3**(4), 49–59.

Sherman, R. (1985). The Averch and Johnson analysis of public utility regulation twenty years later. *Rev. of Ind. Org.*, **2**, 178–93.

Sweeney, J. (2006). The California electricity restructuring, the crisis, and its aftermath. In *Electricity Market Reform: An International Perspective* (F.P. Shioshansi and W. Pfaffenberger eds.). Elsevier.

Williamson, O. (1975). *Markets and Hierarchies: Analysis and Antitrust Implications*. The Free Press.

Williamson, O. (1985). *The Economic Institutions of Capitalism: Firms, Markets, Relational Contracting*. The Free Press.

Wilson, R. (1993). *Nonlinear Pricing*. Oxford University Press.

Wilson, R. (2002). Architecture of power markets. *Econometrica*, **70**, 1299–340.

Zarnikau, J. (this book). Demand participation and demand response: Evidence from US ISOs, Chapter 9.