Interface Between Engineering and Market Operations in Restructured Electricity Systems

HUNG-PO CHAO, SHMUEL S. OREN, FELLOW, IEEE, ALEX PAPALEXOPOULOS, FELLOW, IEEE, DEJAN J. SOBAJIC, SENIOR MEMBER, IEEE, AND ROBERT WILSON

Invited Paper

We examine the impact of wholesale markets on operations of the bulk power system and elaborate some basic implications of engineering practices for designs of wholesale markets. This analysis is intended to provide a basis for enhancements to existing principles of engineering management. Wholesale markets bring economic and financial aspects that alter the context in which system operations are conducted, and introduce incentive and benefit–cost considerations that might alter operating procedures that previously were based on reliability considerations. The principles addressed are those relevant to the interface between engineering aspects of system operations, and economic aspects of market operations. We outline ways that engineering practices developed in the era of vertically integrated utilities might be adapted to the wholesale markets introduced since restructuring began in 1998 in the United States

Keywords*—Electricity restructuring, market design, market operation, power systems operation.*

I. INTRODUCTION

In this paper we concentrate on two important aspects characterizing the restructuring of the electric power industry in the United States. The first is that reliability is now affected substantially by the fact that resources needed for grid operations are purchased from many independent participants in wholesale markets, each with its own profit motive. Thus the reliability obtained from the previous "command and control" system is degraded by a new dependence on market mechanisms that bring their own sources of unreliability and volatility. The effects are both short-term, as in day-ahead and real-time markets for energy, reserves, and congestion relief; and long-term, as in investments in transmission and generation capacity. The severity of these effects is heightened by the fact that in the United States the transmission system operator (TSO)[1] must adhere rigorously to the rules established in its tariff approved by FERC, thus eliminating much of the role for operators' judgment and discretion that was implicit in vertically integrated systems. The second important aspect of restructuring is that engineering standards and practices imply financial consequences for market participants. Because choices of operating standards can severely impact their profits, individually and collectively, market participants inevitably insist that these choices be subjected to comparisons of aggregate benefits and costs. Their insistence is especially strong regarding those costs that the TSO passes to participants via grid management charges. Many of the operating standards and practices inherited from vertically integrated utilities are widely accepted as necessary or desirable, but others are subject to new scrutiny and, to the extent there are tradeoffs at the margin, require justification on economic grounds.

Restructuring at the state level allowed regulated utilities to purchase energy supplies in wholesale markets from independent generators, some of whom acquired their facilities from among the units divested by utilities. The key enabling regulatory initiative at the federal level was FERC's Order 888 in 1996 and later Order 2000 that set standards for a TSO, financially independent of all market participants, with responsibility for grid management over a wide area encompassing several of the control areas previously managed by local utilities. Participating utilities

Manuscript received October 1, 2004; revised June 1, 2005. This work was supported by the Electric Power Research Institute (EPRI).

H.-P. Chao is with the Electric Power Research Institute, Palo Alto, CA 94304 USA (e-mail: hchaO@epri.com).

S. S. Oren is with the IEOR Department, University of California at Berkeley, Berkeley, CA 94720 USA (e-mail: oren@ieor.berkeley.edu).

A. Papalexopoulos is with ECCO International, San Francisco, CA 94104 USA (e-mail: alexp@eccointl.com).

D. J. Sobajic was with the Electric Power Research Institute, Palo Alto, CA 94304 USA. He is now with Grid Engineering LLC, San Jose, CA 95120 USA (e-mail: dsobajic@gridengineering.com).

R. Wilson is with the Graduate School of Business, Stanford University, Stanford, CA 94305 USA (e-mail: rwilson@stanford.edu).

Digital Object Identifier 10.1109/JPROC.2005.857491

[1]We use the term TSO to represent either an independent system operator (ISO) or the system operator of a regional transmission organization (RTO).

retained ownership and maintenance responsibilities for their transmission assets, and remained entitled to cost recovery under state regulations, but the TSO acquired responsibility for ongoing allocation of usage, congestion management, provision of reserves, and all real-time operations such as balancing. Participation was voluntary unless mandated by state regulators, so some utilities exempt from state regulation, such as municipal utilities and rural cooperatives, chose not to participate. All TSO's are nonprofit public-benefit corporations (often derived from previously existing cooperative power pools), although FERC allowed for-profit independent transmission companies (ITCs) to exist within a TSO. The nonprofit status of a TSO requires that the costs it incurs are recouped via charges to market participants.

The key economic requirement of Order 888 is that a TSO must provide open access on nondiscriminatory terms set forth in an open access transmission tariff (OATT) approved by FERC. Initially, FERC allowed wide latitude in the tariffs it approved. Former power pools offered their own energy markets, and used security-constrained economic dispatch. Other TSOs were substantially decentralized; in particular, California relied upon separate power exchanges and allowed participants to do their own unit commitments, scheduling, and provision of reserves. But this latitude dissolved after the 2000–2001 crisis in California, and FERC now prefers a highly centralized design, called the standard market design (SMD), modeled on those of the former power pools (NEISO, NYISO, and PJM Interconnection). The wide latitude allowed in governance of TSOs also dissolved and FERC now requires a Board comprised of independent directors, rather than market participants and/or stakeholders.

FERC's Order 888 and subsequent orders left engineering requirements unchanged. As in previous vertically integrated utilities, procedures recognized as good practice were based on standards established by the North American Electric Reliability Council (NERC) and regional coordinating councils, to which FERC delegated nearly all such matters. The advent of TSOs, however, rendered the previously established NERC operating policies, which are centered on control area operations, to become difficult to apply and enforce. Under the new NERC functional model [1], the traditional control area responsibilities are assigned to the market entities performing functions described in the model.

We begin by examining the implications of current operating practices on market design and operation in Section II, followed by a discussion and examples in Section III of how the new market-based environment impacts system operations. In Section IV we provide a series of general prescriptions for system operations and in Section V we outline commercial grid management practices that address the new challenges faced by the TSOs charged with facilitating efficient market operations while meeting their obligation to provide reliable service through market-based procurement of resources. We do not attempt to provide here a comprehensive survey of engineering management and grid operations nor the principles of market design and operations. Hence we only cite a handful of references that serve as background material for the broad issues and perspective addressed in this paper.

II. Implications of Operating Practices for Markets

The latitude initially allowed by FERC enabled several different market designs to be implemented in the period 1998–2000. Experience from the first five years of these markets showed that it is a difficult task to design an efficient integrated system of forward and spot markets for energy, transmission, and reserves. Nevertheless, there were many lessons learned, and now there is a substantial consensus about most of the major components. We do not elaborate these here, but rather focus on what has been learned about the impact of operating procedures on market design.

A. Price All Scarce Resources

A basic economic principle is that markets can allocate resources efficiently only if they are complete; i.e., prices are established for all scarce resources. In electricity markets, this maxim applies to the basic system resources, such as energy and transmission, and also to whatever additional resources are needed by the operating engineers to sustain reliability, such as reserves and reactive power.

In some cases multiple resources are bundled together, but even so there is a price for the bundle. An important example is locational marginal pricing (LMP). When LMP is used the "nodal price" for energy at a bus is actually the sum of the system-wide market price for energy and the local price for injection or withdrawal from the grid. In particular, the difference between two nodal prices is the congestion charge for transmission between them, reflecting the extent of congestion. The price for transmission reflects the scarcity value of transfer capacity on all the paths between the two nodes (and also the power transfer distribution factors derived from Kirchhoff's Laws). These scarcity values represent the marginal cost of redispatch to alleviate congestion on affected transmission lines. Some TSOs obtain these scarcity values as Lagrange multipliers calculated as part of an optimal power flow (OPF) calculation. In the decentralized California system they were obtained directly from the difference between the last pair of bids for incremental and decremental energy redispatch that the TSO accepted in order to alleviate congestion on a zonal interface—which sufficed, since there were no loop flows among its zones.

California's system of zonal pricing of energy provided strong evidence about the potentially severe consequences of not pricing all scarce resources. The TSO did not charge for intrazonal congestion in its day-ahead market, and then in real-time paid for increments and decrements to alleviate congestion. Therefore, a participant could game the system by overscheduling transmission usage day-ahead and then in real-time be paid to alleviate congestion that it had caused. A sure cure for this "DEC game" is to charge for congestion on all lines, through nodal pricing, flowgate pricing on impacted lines or both (see [2] and [3]).

All TSOs tolerate unpriced resources to some degree in order to simplify their markets. But if these resources be-

come scarce then it becomes necessary to adopt some form of efficient rationing, which in a market means establishing prices that reduce demand or expand supply. For example, most TSOs do not pay for reactive power but as noted in a recent FERC report [4], when reactive power is scarce a positive price would attract new supplies from generators paid to adjust their reactive output or induce consumers to purchase equipment that reduces their reactive loads.

Some resources are priced imperfectly for practical reasons. For instance, some TSOs use a linear pricing scheme to charge for losses even though losses are generally a quadratic function of transfers. In the case of unit commitments needed to assure reliability, several TSOs simply assure the supplier that the cost of start-up and minimum energy will be recouped. Generation needed for local reliability or voltage support is treated similarly in that a so-called reliability-must-run (RMR) contract reimburses the supplier's incremental cost.

An important extension of the "price all scarce resources" principle is that scarce resources must be identified. An example is the range of "products" specified in the day-ahead market for ancillary services. A TSO that uses regulation excessively to cope with steep ramps in the morning and evening might benefit from purchasing an ancillary service intended specifically for load following. Identifying what it is that is scarce is a basic problem in the design of markets for ancillary services. From an engineering viewpoint, the TSO values reserves largely in proportion to the speed of response; i.e., the start-up interval (if any) and the unit's ramp rate. Thus, the standard categories of operating reserves (regulation, spin, nonspin, replacement) reflect rough approximations of response times, even though an ideal market might pay a reserve unit directly for its ramp rate. An implication for market design is that prices should reflect downward substitution, in the sense that an unsuccessful bid for spin (for example) should be allowed to compete in the market for nonspin reserve capacity even if the unit will actually be spinning and synchronized. However, when used for nonspinning reserves the unit should not be paid more than spinning reserves (a phenomenon known as price reversal) since that might create incentive for misrepresentation of capability.

Ignoring the role of a valuable attribute such as a quick start or a fast-ramp rate leads to inaccurate pricing that can worsen the situation. For instance, a TSO that ignores the advantages of CTs in its day-ahead optimization of unit commitments and scheduling can find itself with insufficient fast-response units if these mobile units are removed when the owners perceive better profits elsewhere.

Engineering requirements for local reliability pose special difficulties. Restructured wholesale markets for energy are designed specifically to reap the gains from trade over large regions, subject to transmission constraints. But within a trading region there are usually urban areas with localized needs for voltage support or for sufficient generation on-line to provide security against outages. To some extent these local "nomogram" constraints can be included within the set of transmission constraints, and therefore their implicit prices are reflected by the nodal prices in these areas. But engineering procedures typically require the stronger controls provided by RMR contracts, and further, generators in such an area often have substantial market power, so these contracts provide remuneration based on incremental cost. Thus reliance on markets is replaced by direct controls and cost-based remuneration to ensure local reliability. Similar situations occur when constraint violations are not amenable to market solutions due to insufficient independent bids or lack of independent resources that can resolve the violations. In such cases the operator typically employs "out of market" (OOM) dispatch procedures using nonmarket settlement rules to compensate the used resources. In ERCOT, for instance, a constraint violation is considered to have a market solution only if there are at least three independent resources, none of which is pivotal (i.e., absolutely needed), that can resolve the violation.

Another example in which the role of markets is curtailed occurs in the real-time balancing market where TSOs impose penalties for excessive uninstructed deviations from schedules set day-ahead. Penalties discourage under- and overscheduling in the day-ahead markets, and also discourage "price chasing" in real time, either of which magnifies the operating engineers' difficulties in maintaining reliability.

B. Market Implications of Physical Feasibility

Physical feasibility is an engineering requirement that has a major impact on the design and performance of markets. Feasibility in real-time is necessary, of course, but the main impact on markets stems from the engineering standard that requires physical feasibility in day-ahead markets.[2]

FERC has insisted on this standard since the 1999–2001 episodes in California in which huge imbalances in the real-time market caused severe problems for the TSO. Initially there were substantial imbalances because loads did not match supplies from RMR contracts. Later, demand imbalances stemmed from massive day-ahead underscheduling by utilities, compounded by gaming of its zonal pricing system, "megawatt laundering," "phantom schedules," and other strategies employed by market participants. The prevalence of these strategies, and their severe effects on system reliability, showed that power systems cannot rely on individual market participants to ensure overall physical feasibility; indeed, the clear conclusion is that financial incentives and gaming opportunities can easily thwart the engineers' attempts to maintain reliable operations.

These dislocations in the markets reinforced the engineering viewpoint that physical feasibility should be ensured in forward markets just as in real-time operations. The net effect is that, for practical purposes, the day-ahead market is the final market for trading on a large scale. Subsequent deviations are addressed using resources from the balancing market and from reserves, and further, the volume of transactions in the balancing market is kept small by imposing

[2]Some TSOs (e.g., in Australia) use only a single settlement at real-time prices, so the day-ahead market outcome is only indicative, not binding financially. The U.K.'s TSO receives schedules only a few hours before real-time; thus, there is no prior enforcement of physical feasibility.

penalties on uninstructed deviations outside a small band, say 5%.

Day-ahead physical feasibility has two components. The first is that for each scheduling period (say, hour) of the next day, scheduled supply and demand balance, adequate reserve capacity is provided, and all transmission constraints are satisfied. This standard is consistent with the view that the day-ahead market is essentially the final market, even if the balancing market facilitates a small amount of subsequent trading in real-time based on an optimized power flow (OPF). The second component is the one that produces tension between engineering and market considerations. Especially in those systems derived from former power pools, the TSO schedules further supply-side resources through the so-called reliability unit commitment (RUC) to meet the load it predicts, rather than the load implied by demand-side schedules. In its SMD, FERC endorses these interventions by the TSO.

There are two important implications for the design and performance of markets. First, multiple settlements are needed: day-ahead transactions are settled at the day-ahead prices, and then deviations are settled separately at the real-time prices. This creates a strong incentive for market participants to arbitrage the difference between the day-ahead and real-time prices. This motive is suppressed, however, by significant penalties for uninstructed deviations from day-ahead schedules. An exception occurs in systems like PJM that allow "virtual bids" in the day-ahead market: these bids are explicitly identified as purely financial bids (i.e., not backed by physical resources) intended for arbitrage and covered financially only by the opportunity to close out the position in the real-time market. The TSO retains the right to reject virtual bids that could jeopardize physical feasibility in the balancing market. The TSO can also revoke settlements of point-to-point financial transmission rights (FTRs) for "phantom congestion" created by virtual bids. The distinction between those bids that are backed by physical resources and those that are not reflects the fundamental character of power markets; specifically, correcting large imbalances in the real-time market can be costly, and more important, can threaten reliability.

The second implication is that the TSO's interventions can have large distributional effects on market participants. By scheduling additional supply-side resources to cover a positive difference between predicted and scheduled loads, the TSO lowers real-time prices. This role accords well with prudent engineering practice, but distorts price signals that may affect investment incentives, as we discuss later. The next subsection discusses further distributional effects resulting from the TSO's RUC.

C. Distributional Effects of Engineering Practices

The initial market design in California differed significantly from those in the northeast derived from former power pools in that transactions in California were settled at market-clearing prices that equated demand and supply. FERC's SMD rejects this design in favor of optimized unit commitment and scheduling by the TSO, with transaction prices derived from shadow prices (Lagrange multipliers) on binding constraints. This optimization is comprehensive, including balancing of energy supply and demand, eliminating transmission congestion, and scheduling reserve capacity. From an engineering perspective, an overall optimization is the surest way of ensuring physical feasibility and reliability. The effects on prices paid or received by market participants can be significant, however. Here we enumerate several such effects to illustrate problems with which the market design must cope.

1) Transmission Congestion Charges: In an LMP system the nodal prices reflect the least cost of serving an incremental unit of load at the node while observing all the active transmission constraints. When a resource is active and available at the node, the LMP price equals the marginal price of the resource. However, the LMP price at a pure load node can exceed the marginal price of any of the active resources. For example, if serving an additional MWh at node A requires incrementing a \$40/MWh generator by 10 MW and decrementing a \$30/MWh by 9 MW in order not to violate transmission constraints, then the nodal price at node A equals the incremental cost of \$130/MWh. The congestion charge is the difference between the nodal prices at the withdrawal and injection buses, which can be positive or negative. This difference reflects the opportunity cost of selling the power at the injection node and buying it back at the withdrawal node at the respective nodal prices. The participant pays this difference for every MWh transferred between the two buses.

An important aspect is that this payment need not bear any particular relationship to the actual average cost of redispatch. For example, suppose that in a given hour the excess demand is 100 MWh on a line with capacity 1000 MWh, and the net average and marginal costs of redispatch are \$5 and \$10 per MWh; then participants pay $1000 \times \$10 = \$10\,000$ in congestion charges, whereas the actual cost of redispatch is $100 \times \$5 = \$500$. This discrepancy between what participants pay in congestion charges and the actual cost of eliminating congestion has profound implications for the market design. Because customers of load serving entities (LSEs) have historically paid for transmission assets, the proposed solution in the SMD is to assign congestion revenue rights (CRRs) to LSEs (rather than to transmission owners as previously) in proportion to their historical usage and predicted load growth and thus largely immunize them against congestion charges. CRRs are not necessarily perfect financial hedges, however, because the assignments are limited by a joint feasibility requirement, and in some versions they entail an obligation to pay when the nodal difference is negative.

The basic feature that congestion charges can differ greatly from actual costs of redispatch is an instance of a recurrent theme in the subsequent examples. That is, prices based on the marginal costs of correcting a problem have large distributional effects on those participants using the system after the problem is solved.

2) Cooptimization of Energy and Reserves: From an engineering viewpoint it seems prudent to schedule sufficient

capacity to meet the sum of the predicted load and the reserve requirement. The capacity held in reserve is then paid its opportunity cost (i.e., the difference between the resulting market price of energy and the unit's marginal cost as bid) since this is presumably what it could earn as profit were it not required to be available as reserve capacity. From an LSE's perspective, however, this practice raises the energy price above what it would be if the reserve capacity were not treated as though it were supplying energy. And from the perspective of a generation unit near the margin, it creates an incentive to underbid so that it is paid a larger opportunity cost if it is assigned to reserve. As with congestion charges, cooptimization is an example of a practice that solves the immediate problem (scheduling sufficient reserve capacity) via a procedure that introduces price effects elsewhere; in this case, by raising the price of energy even though reserve capacity is not scheduled to produce energy. Other designs avoid this problem by conducting a separate market for reserve capacity in which a generation unit offers a bid for capacity availability—together with an indication of its marginal cost that is used in real-time to determine its position in the merit order for calling reserves.

In some cases participants have insisted on procedural rules for the balancing market that have severe effects on real-time energy prices. In Texas' ERCOT system, the TSO is prohibited from calling reserves until all bids in the balancing market are exhausted and when called the energy produced by reserves is paid the market-clearing price of balancing energy. This requirement can drive the energy price to exorbitant levels, exacerbated by the induced incentive for suppliers to always offer a very high price for a small quantity in the balancing market—so-called hockey-stick bidding. At the Mid West ISO (MISO), on the other hand, reserves are only called to produce energy under extreme circumstances and the market-clearing price for balancing energy is automatically set to the price cap whenever such an event occurs. In California, energy from reserves is added to the balancing energy stack and dispatched in merit order until the reserve level drops to 6%, at which point the reserves' energy offers are skipped. A better rule displaces reserves to later positions in the merit order by applying an "adder" that reflects the added cost of reserve deployment and the scarcity of fast-response resources. The adder can be interpreted as either the explicit cost of activating replacement reserves, or the implicit impact on reliability from reducing reserve capacity.

3) Unit Commitment: There are two major options for scheduling in a day-ahead market: self-commitment or centralized unit commitment. With self-commitment, generators are responsible for their own commitment, i.e., the decision of which generating units will run the next day. By contrast, with centralized unit commitment, these decisions are taken optimally by the TSO for all generating units based on their cost/bid information and their technical characteristics. The greatest disadvantage of self-commitment is its inconsistency with the must-offer obligation. Must-offer obligations have been increasingly used in the United States after the California energy crisis. A resource that is subject to the must-offer obligation must bid in every hour when it is available. Its only mechanism to incorporate technical constraints and internalize start-up and no-load costs is through the offer prices. The bidding process provides bidders the needed flexibility to internalize their constraints in their offers. Extreme offer prices, however, are problematic. High offer prices may interfere with market power mitigation mechanisms and market monitoring activities. Low offer prices, on the other hand, may interfere with regulatory limitations to protect against predatory pricing. Many market designs prohibit bids below the relevant variable cost. Self-commitment does not appear to be a viable option for the day-ahead market with must-offer obligations, as it would create large imbalances, considerable burden for operators during the balancing process, and great risk to generators, which could become a barrier for entry.

In some markets that use self-commitment (e.g., ERCOT), the scheduling-feasibility risk is mitigated by allowing portfolio bidding. With portfolio bidding, generators submit aggregate offers for the portfolio of their generating units, and then specify the unit-specific schedules that comprise the accepted portfolio offer after the latter clears the day-ahead market. Portfolio bidding, however, is clearly not a good choice in market with one or two dominant players since it would give them a tremendous competitive edge compared to new entrants. Furthermore, experience in ERCOT shows that lack of unit-specific offers results in inefficient resource use and in cumbersome settlement procedures.

In centralized markets, unit commitments selected by the TSO to ensure adequate reserve capacity can occasionally have severe effects on energy prices. The standard procedure is to select units according to the merit order based on the least total cost of start-up and minimum energy generation. This procedure favors selection of units such as CTs that have low start-up costs and low minimum generation levels, but typically these units also have high marginal costs for energy generation. Consequently, if the selected units are included in the construction of the aggregate supply curve then the resulting price of energy must be at least as high as the highest marginal cost among the selected units. This is another instance of a sensible engineering procedure for solving a reliability problem that can distort energy prices.

In the terminology of economics, these are all examples of pecuniary externalities. Measures undertaken to solve a reliability problem can distort the prices paid and received by demanders and suppliers of energy. Such payments from demanders to suppliers of energy are "merely" financial transfers of funds among participants (i.e., they are pecuniary transfers) that have no particular implications for overall productive efficiency of the market, yet they have potentially very large effects on the division of the gains from trade between buyers and sellers. Moreover the incentive effects can ultimately jeopardize full realization of the potential gains.

D. Efficiency Effects of Engineering Standards

We mention here briefly the effects of engineering standards regarding the allowed variation of such power quality

parameters as frequency, voltage, etc. Such standards are necessary for stable operation of a large interconnection, and for retail customers they are valuable attributes of service. Except for risks of grid collapse, much of the engineering effort regarding reliability focuses on maintaining these standards, and therefore a portion of the cost of grid management can be attributed to these standards. Unfortunately, these standards have not been subjected to a benefit–cost analysis in the years since wholesale markets were restructured. One of the fundamental premises of the restructuring was that market forces will determine the socially desirable level of reliability and will facilitate customer choice between reliability and price. This idealized market concept along with the promised efficiency gains it entails has been abandoned in virtually all restructured systems in favor of maintaining the traditional generation adequacy standard based on a tolerable outage rate that translates to a planning reserve margin requirement. Arguably, setting a resource adequacy standard rather than allowing market forces and demand response to set these levels, forfeits much of the long-term efficiency gains from restructuring. However, even with centrally set standards there is still the pending question of how to set investment targets and incent private investment through price signals.

E. Long-Term Planning and Resource Adequacy

We conclude this section with a discussion of the planning and investment problems resulting from restructuring. Previously, utilities undertook integrated resource planning that coordinated investments in generation and transmission to meet predicted load growth. This role has largely evaporated because independent power producers now undertake most investments in generation. Coordination is now more difficult because transmission expansion must be based on predictions about the location and magnitude of new generation units; similarly, investors in generation must rely on predictions about investments in grid resources. The TSO can facilitate coordination by proposing a long-term plan for expansion of the grid, thus providing investors in transmission and generation with shared expectations about future developments. In fact, however, reliance is placed mainly on market mechanisms.

A common view is that LMP is the appropriate market mechanism. Persistently high nodal prices in an area presumably signal profitable opportunities for investments in new generation there. Similarly, high congestion charges across a line indicate the desirability of expanding its transmission capacity. In fact, however, these volatile short-term price signals are imperfect indicators for long-term investments. One adverse factor is that an investment on a large scale eliminates the profit opportunities by depressing the nodal price or the congestion charge. In theory, transmission capacity might be adjusted so that congestion rents exactly cover the incremental capacity cost of transmission expansion and match the incremental cost of relieving congestion through out-of-merit dispatch. Unfortunately that naive theoretical paradigm is invalidated by the lumpy nature of transmission investments. Further, transmission investments typically have external effects throughout the grid, so an investor who expands the capacity of one line need not capture all the benefits created. Transmission investments are also a source of controversy, since the reduction in energy prices achieved at one end is offset by an increase in prices at the other end. Controversy also pervades the choice between transmission and generation; e.g., generators argue that they should have the first opportunity to build new capacity in a "load pocket," rather than relying on expanded transmission to solve the problem—an argument that has added force due to the longer lead times for new transmission compared to new generation.

The deficiencies of nodal prices as signals for new investments compounded by various intervention schemes designed to mitigate the exercise of market power by generators, have led many TSOs in the United States and abroad to rely on artificial products and market mechanisms to stimulate investment. Following the examples of the eastern TSOs, FERC's SMD endorses the use of markets for installed or available capacity (ICAP) based on requirements imposed by the TSO. In such a system, each LSE must provide evidence to the TSO that it owns or has contracts or claims for sufficient capacity to cover its peak load plus a reserve margin. Those LSEs and generators without sufficient contract cover can trade claims on generation capacity in a monthly market, thus enabling the LSEs to provide the required evidence to the TSO. LSEs that are short face a penalty and the TSO procures the missing capacity as provider of last resort. ICAP markets are "bipolar" in nature due to the fact that within the short time frame of a month the supply and demand are both highly inelastic so the clearing price is either near zero or near the level of the shortfall penalty. Furthermore, the ICAP markets failed to meet the reliability goals of the TSOs due to the minimal deliverability requirements imposed on the ICAP providers. Recent reform proposals for capacity markets by the NYISO and NEISO have added a locational dimension to the ICAP and an administrative demand function that adjusts the price smoothly when supply falls short or exceeds the target quantity. Nevertheless, the proposed locational installed capacity mechanism (LICAP) [5] in New England is highly controversial and is being contested by the New England states' authorities due to the high estimated financial impacts on consumers. It is yet unclear whether revenues from sales of capacity claims in an ICAP or LICAP market might suffice to stimulate new investment, but perhaps they may succeed in retaining obsolete or inefficient generation units that would otherwise be shut down.

The design of mechanisms to incent investment can be materially improved by changing the product traded from capacity claims, which are merely paper "chits" showing that the capacity exists, to actual claims on energy output in the form of fixed-price contracts, or preferably, call options with physical cover [6]. It is also essential that the traded contracts be sufficiently forward looking so as to allow new entrants to participate and contest the prices offered by the incumbent generators. Such products may be self-provided through bilateral supply contracts and curtailable loads. As demand

response increases and the markets for financial hedges matures, the role of the TSO in administering a capacity mechanism may be reduced to a monitoring role. We discuss this alternative approach in more detail later.

III. Implications of Restructured Markets for Engineering Principles

We now turn to an examination of the various ways that the new role of wholesale markets might affect the engineering principles developed in the era of vertically integrated utilities. We concentrate on two key features. One is that engineering operations now depend on procuring needed resources from markets. The second is that management of the grid, and management of those markets that the TSO conducts, are now required to support and enhance the efficiency of trading among participants.

Power systems are built and operated to supply customers with electrical energy at least total cost. The costs include both the supply-side capital and operating costs and the demand-side costs to customers of inevitable imperfections of power quality and reliability. Power quality for customers includes attributes such as frequency and voltage that are also relevant for stable operation of the grid. Reliability includes two components:

- resource adequacy to ensure that sufficient facilities exist within the system to satisfy the demands from loads i.e., the ability to provide continuous service under stable conditions;
- system security (operational reliability) to ensure that the system can recover from disturbances, i.e., the ability to withstand perturbations.

Restructuring has brought to the fore a tension between the incentives of engineers and market participants. In the United States, FERC requires that each TSO is nonprofit, and thus passes its costs to customers via uplift charges, and its governing board cannot include market participants.[3] This requirement brings the risk that the TSO acts bureaucratically, incurring whatever costs makes its managerial task easier, subject of course to the rules in its tariff. In terms of engineering practices, the chief impediment to a "gold-plated" system, in which power quality and system reliability are enhanced regardless of cost, is the resulting suppression of investors' incentives to build new transmission assets.

A. New Problems Raised by Reliance on Markets

As emphasized previously, reliability is now affected substantially by the TSO's dependence on wholesale markets to obtain resources needed for grid operations. Participants in wholesale markets have their own profit motives that can diverge from engineering requirements for reliability of the grid. The reliability obtained from the previous command-and-control system is now degraded by reliance on markets that bring their own sources of unreliability and volatility. And the world of markets is far different than a power system—financial incentives play the role of physical laws, and gaming strategies play the role of grid congestion or voltage collapse. Especially frustrating for operating engineers is that command-and-control is partly replaced by dependence on others' investments and the offers that they submit in markets. This dependency is also passive to the extent that the TSO must follow the rules in its FERC-approved tariff. The role of operators' judgment and discretion in vertically integrated systems is now severely constrained.

The conflicts and risks now evident in restructured wholesale markets stem in part from lack of an integrated methodology for grid management. System operators must now consider commercial strategies for dealing with commercial risks. For example, it is now common to settle forward transactions at day-ahead prices, and then to settle real-time deviations at different prices. This practice makes day-ahead transactions financially binding, even though they are not completely binding in terms of the physical commitments relevant for engineering operations. Other examples that have major impacts on operating procedures are decisions about whether to contract long-term or day-ahead for options on reserve capacity, for RMR generation used in voltage support, and for services such as regulation, load following, and balancing energy. Operating procedures also differ depending on whether:

- reserves are obtained day-ahead in separate markets or in a consolidated market for energy, reserves, and transmission;
- the consolidated market includes unit commitment and accounts for startup and ramping constraints;
- participation in the TSO's energy and reserve markets is voluntary or mandatory;
- the extent to which bilateral contracts are exempt from ongoing dispatch control;
- restrictive market protocols (e.g., the previously cited example of the rule in ERCOT that requires exhaustion of balancing bids before reserves are called).

Operating engineers must now support the wholesale market while also relying on it for critical resources. This reliance poses fundamental problems because grid management now depends on the following.

- Distributed control, since generating companies are autonomous profit maximizing agents.
- Indirect control, since procurement of resources occurs through price-mediated markets.
- Imperfect control, since market participants can renege on deals arranged in markets (subject only to financial penalties), and tariff provisions limit the system operator's discretion.
- Noisy control, since the TSO has imperfect information about generators, and participants' responses to dispatch directives are affected by incentives and commercial considerations outside the TSO's purview.
- Complex control, since a TSO manages a regional system with multiple control areas, thousands of buses and lines, hundreds of market participants, and a sequence of forward and spot markets for energy,

[3]Those TSOs derived from power pools provide somewhat stronger controls. Matters are substantially different in countries such as the U.K. and Sweden, in which an independent transmission company (ITC) both owns and operates the grid, and under performance-based regulation, has strong incentives to minimize the uplift charges to customers.

reserves, and transmission. New resources such as dispatchable loads, intermittent energy sources, and distributed generation add further complexity.

Increasingly, these problems are eased by technical developments such as more reliable communication and real-time monitoring, more complete state estimators, and frequent reoptimization of dispatch. Nevertheless, implementation of good engineering practices depends on close coordination with the design of markets for grid resources and the specification of procedural rules and contract forms, and it must be done within the many constraints imposed by tariff provisions.

We conclude this section with some examples that illustrate the many ways that commercial considerations now infuse grid management.

Market Power. Because generation units in areas with local reliability problems can have substantial market power, some TSOs find it necessary to impose reliability-must-run contracts on such units or to override the market by invoking "out of market" dispatch and settlement procedures.

Surplus Resources. Seemingly excessive transmission capacity is now justified by the goal of suppressing the market power of generation companies. TSOs have similar motives to impose installed or available capacity requirements that exceed reliability requirements. An investment criterion that counts only the net benefits to customers often replaces the traditional social welfare maximization criterion.

Uncoordinated Resources. Vertically integrated utilities tried to maintain an optimal mix of remote and local generation, recognizing that redispatch is a substitute for transmission. The TSO still uses redispatch to eliminate congestion, but the effects on investment incentives are greatly altered. As described earlier, the TSO's congestion charges are based on the marginal cost of redispatch, and then applied to the actual flow rather than the curtailed flow. Hence there is no single entity in the system whose incentives encourage optimal substitution between generation and transmission. This reflects the fact that restructuring abandoned the integrated resource planning process previously used by utilities.

Insufficient Local Resources. Complaining that payments for reserve status were insufficient, some owners of CTs removed them from the New England system, leaving the TSO dependent on high offers from hydro resources to meet its largest contingency, failure of the transmission line from Quebec. The TSO recently sought to attract CTs via a special procurement auction. A pervasive problem is maintenance of local reliability in urban areas when there are few local generators and siting of new construction is difficult and often long delayed. TSOs have imposed reliability-must-run contracts on key generators to deal with these problems.

Dependence on Remote Resources. Restructured markets encourage energy trading over large regions. FERC's Order 2000 established guidelines for formation of regional system operators (RTOs), but this initiative stalled with the April 2003 White Paper in which FERC yielded to the reluctance of many states to join such organizations. The likely prospect is that the scope of system operators will not extend as wide as the markets they support. Thus commercial decisions by entities outside a TSO's control areas will continue to be a major source of risk, and "seams issues" will continue to be significant factors. The procedures of some TSOs can jeopardize others, as in the case of PJM, which requires participants to recall exports when the TSO declares an emergency condition. The fact that many TSOs depend on imports to meet their loads is a basic threat to reliability.

Unscheduled Maintenance. The frequent problems caused by unscheduled, and sometimes opportunistic, maintenance of generation units in the first years after restructuring led TSOs, encouraged by FERC, to include control and scheduling of maintenance as an integral part of grid management.

Market Failure. The California ISO found it necessary to impose a punitive default congestion charge ($250/MWh) when occasionally there were insufficient bids for incremental or decremental adjustments to eliminate congestion on a zonal interface. The economic analysis of markets rarely accounts for market failure, but from an engineering perspective it is absolutely necessary to develop operational procedures to cope with "bid insufficiency" in the markets for the resources that are needed for reliability.

These examples illustrate that a TSO now encounters a variety of operational problems that stem from commercial considerations in the markets that it is now obligated to support. An implication is that engineering principles must be extended to encompass the relevant commercial considerations, and operational procedures for managing the grid must include commercial strategies for dealing with commercial problems.

IV. Prescriptions for Engineering Practice

In the following subsections we specify several prescriptions that elaborate in more detail the engineering aspects of general principles. These summarize our current thinking about the implications of wholesale markets for engineering standards and procedures. Our intention is to initiate discussions about how existing engineering principles might take account of the new problems that operators face when they must support wholesale markets, and also rely on them for resources. Thus these prescriptions are intended to suggest concrete ways in which engineering principles can be implemented in a market context.

A. Identify Scarce Resources and Valuable Quality Attributes

From an economic perspective, the "DEC game" in California revealed that its zonal pricing system was vulnerable to gaming because: 1) intrazonal congestion was not addressed in the day-ahead market, and thus 2) those participants who caused congestion could profit from being paid for redispatch to alleviate it. From an engineering perspective, it is item 1) above that is instructive. Markets cannot allocate resources efficiently when some scarce resources are not recognized, and experience has shown vividly that power markets are especially vulnerable to gaming that exploits such

deficiencies. Now that power systems rely on markets, an important engineering task is to identify explicitly the scarce resources. This is the first step in designing mechanisms to allocate these resources efficiently. The mechanism adopted might be a market, but depending on circumstances it might be some other rationing scheme, such as assignment (or sale) of priorities. The choice of allocation mechanism is partly an economic problem, but to the extent that reliable operation requires tight control, engineering considerations might dictate another rationing procedure. Although the market design itself might be outside the application of engineering principles, an efficient game-proof design is impossible unless the resources to be allocated are clearly identified and the technology supports the distinction between products traded in the market.

An example of an application of this prescription is the specification of "nomogram" constraints. TSOs increasingly recognize that their procedures for maintaining reliability in local areas include specific constraints on generation, transmission, and local reserve capacity. Like N-1 security constraints, these can be included in the formulation of a security-constrained economic dispatch, and thus reflected in the resulting nodal prices for energy.

A second example illustrates an application to quality attributes. Those systems with a large "first contingency" that must be covered by reserves increasingly recognize that a large supply of fast-start, fast-ramp reserves is needed. Identifying this need is the first step toward establishing a market mechanism that rewards generation sources that provide this capability. This example can be extended to the more general task of identifying the valuable quality attributes of reserves; if a fast-ramp rate is valued by the engineers then the market for reserves can be designed to pay for, and thus attract, capacity with this attribute. Similarly, if reactive power becomes scarce then a market that establishes prices for reactive power may be the best way to attract additional supplies and load response beyond those obtained from cost-based remuneration.

B. Consider the Incentive Effects of Operational Procedures

A TSO's choice of an operational protocol usually allows some latitude. The specific choice might be dictated by engineering necessity, but if different procedures are cost-effective then the impact on incentives in the markets can be the decisive criterion. We illustrate with several examples.

We mention first an example to the contrary. Engineers appreciate that an overall optimization of energy, transmission, and reserves, subject to all known constraints, has the best chance of minimizing the actual cost of serving load. Such a cooptimization of all markets is now endorsed by FERC. This is an example in which the economist's preference for separate markets for unbundled "products" (energy, transmission, reserves) is overridden in favor of an integrated system optimization, based on recognition that the engineers were right in emphasizing that the various products are too interrelated for separate poorly coordinated markets to work well. These deficiencies of separate markets are partly due to ill-defined products; e.g., the decentralized design first adopted could not take account of constraints on generators' ramp rates and other constraints, and most important, the day-ahead markets were not able to ensure joint feasibility of the units' schedules within the transmission and other operating constraints.

From an engineering viewpoint it may be immaterial in the balancing market whether incremental and decremental bids whose prices overlap are accepted and the units redispatched. From the viewpoint of market participants, however, failure to redispatch based on an OPF forgoes profit opportunities, and from a system perspective causes an inefficiency because not all the gains from trade are realized. Until 2002, California's system did not redispatch overlapping bids (but settled transactions as though they were), and actually the resulting inefficiencies may have been minor compared to the gaming strategies this procedure engendered.

Again, from an engineering viewpoint it may be immaterial which additional units are committed to ensure sufficient reserves at least total cost, but the resulting effects on energy prices can have significant consequences for market participants. Such an optimization of residual unit commitment tends to have many near-optima, and some of these select units with low start-up costs and high energy costs, which then results in a high energy price. More generally, the engineering practice of scheduling additional units to meet the TSO's predicted load, rather than the load scheduled in the day-ahead market reflects an engineering preference for day-ahead assurance of reliability regardless of its impact on day-ahead and real-time energy prices. Most TSOs now give deference to the engineering perspective on this issue, even though no studies have done about the extent of surplus unit commitments that result and their impact.

An important example of an engineering preference based on greater assurance of reliability is the tendency to favor dispatchable generation (versus intermittent sources such as wind, solar, biomass, etc.), and similarly, for reserves to favor generation sources over dispatchable loads. In some cases the preference derives from operators' greater confidence that generation can be counted on to be there when needed. This is a case in which a benefit–cost analysis is relevant. The social benefits of renewable power from renewable energy are significant, and the ability of curtailable and interruptible loads to suppress price spikes is well documented. Therefore, it is important to compare these benefits to the extra costs imposed by nondispatchable generation and dispatchable loads. Only if these extra costs are measured can they be assessed against the bidders who offer them in the market in competition with the usual generation sources.

C. Design the Tariff Rules Carefully

TSOs in other countries have considerable discretion about how they manage the grid, provided they adhere to basic rules (e.g., the Balancing Code in the UK) assuring open access and nondiscriminatory terms. In the United States, however, a TSO must adhere rigorously to very detailed provisions in its FERC-approved OATT tariff. From the first years of restructured markets, two major lessons have been learned. First, rules in the tariff often interfere

with operations; and second, with few exceptions a market participant will use every opportunity to exploit the rules to its advantage, even to the extent of jeopardizing the system's reliability. These lessons reflect the fact that a TSO's tariff is an amalgam of engineering and market rules, and moreover, it is inflexible (FERC approval or rejection of a proposed amendment can take months). A TSO's operating engineers therefore need to be closely involved in the formulation of its tariff. We mention a few of the many examples that might be cited.

An illustrative example concerns the merit order of bids used to eliminate transmission congestion. In its first years the TSO was obligated by its tariff to use such bids in merit order, interpreted as the energy price offered. But the more sensible engineering viewpoint eventually prevailed, namely that the offer price must be divided by the power transfer distribution factor (PTDF) from the injection point to the congested line to obtain a correct measure of the cost-effectiveness of each bid. This is an instance where a simplistic economic view that was contrary to basic principles, written into the original tariff, required years to undo and in the meantime hindered the operating engineers. This example is actually obsolete, since TSOs now use OPFs and other sophisticated software that includes a detailed model of the grid.

Another example, introduced above, concerns markets for ancillary services. Increasingly, TSOs optimize their procurements of operating reserves by allowing downward substitution from a high quality to lower qualities. Thus, a bid rejected for, say, spinning reserve might still compete to supply nonspinning or replacement reserve. The inflexible definitions of ancillary service "products" in the TSOs' initial tariffs reflected an economic prejudice that an engineering perspective would have corrected quickly.

Engineers must sometimes argue forcefully to remove tariff provisions that interfere with operations. A general implication of this prescription is that the United States needs to restructure tariffs so that a TSO's engineers can use again the judgment and discretion they exercised before restructuring. However, the discretion and latitude that is warranted to manage the grid reliably should be exercised with caution so that market solutions and due process are not compromised. An engineering perspective can contribute to the task of separating the economic and regulatory aspects of a TSO's tariff from those aspects that pertain to engineering operations, and excluding provisions that either interfere with operations or attract gaming strategies that can jeopardize reliability. At the very least the tariffs should be rewritten to accurately reflect technical facts about the grid.

V. Commercial Strategies for Grid Management

Long-term bilateral contracting for energy is outside the mandate of a TSO. According to the FERC-recommended SMD, however, a TSO conducts markets that include everything else. The TSO's day-ahead and real-time markets are comprehensive and fully integrated into a single overall optimization of schedules, including energy, transmission, and reserves—and supplementary unit commitments as well. Additional markets are conducted monthly for financial hedges against congestion charges, and for capacity claims that LSEs need to meet their TSO-imposed obligations for installed or available capacity. Participation in the TSO's markets is voluntary, but based on the experience in PJM it is common for 20%–40% of available capacity to be offered for scheduling in the TSO's day-ahead market.

The TSO must deal with many commercial problems that stem from its responsibilities for conducting markets. But here we focus on those aspects of engineering management of the grid that are now infused with commercial considerations. We emphasize that operating practices need to be updated to take best advantage of possibilities for commercial strategies to procure needed resources at least cost. Other countries are far ahead in this respect, which is explained below as we address some specific examples.

A. Reserves

Before restructuring, a vertically integrated utility typically owned and operated sufficient generation capacity that its engineers could each day commit and schedule sufficient units to provide reserves for the next day. This practice is reflected now in reliance on day-ahead cooptimization of units' schedules for energy generation and reserve capacity, or a separate day-ahead market for reserve capacity, for those units not obligated by bilateral contracts. The commercial implication is that the TSO pays prices that are highly volatile.

In some other countries the TSO avoids this volatility by purchasing annual or monthly options on reserve capacity. Those TSOs that are ITCs under performance-based regulation pursue a more general policy. Long-term options are purchased for reserves and also for incremental and decremental adjustments to eliminate congestion, and for balancing energy in real-time. In the United States, FERC's Orders preclude a TSO from direct participation in energy markets, and in particular the nonprofit status of a TSO discourages it from taking long or short positions that expose it to financial risks. But several TSOs have recognized the advantages of long-term options on reserve capacity. For instance, in early 2000 the California ISO purchased options on curtailable loads that substantially eased the crisis during the following year; and as mentioned previously, the New England ISO has conducted a procurement auction to obtain CTs as reserve capacity.

Options on reserve capacity are already used by TSOs in the form of RMR contracts for capacity to meet local reliability requirements, as well as black-start capacity, etc. Like an option, an RMR contract enables the TSO to call on the unit for reserve capacity or energy generation, and if the unit is called then the TSO pays the unit's incremental cost on an annual basis.

The wider use of long-term options on reserve capacity could ease the TSO's dependence on bids offered in daily markets, enhance its control over dispatch, and reduce the volatility of the prices paid. The nonprofit status of a TSO makes it more problematic whether options should be used

for residual unit commitments, adjustments to ease congestion, and/or balancing energy. But these financial and commercial considerations might be balanced against the evident advantages from an engineering viewpoint of having assured long-term supplies available and callable at a cost that is negotiated annually or monthly.

B. Resource Adequacy Requirements

From an engineering perspective it may seem sufficient simply to require each LSE to demonstrate the existence of adequate capacity to cover its peak load plus reserves over a planning horizon. But the commercial implications of this policy are significant, and they affect the extent to which the policy succeeds in promoting its intended goal of stimulating investment. We mentioned previously some of the shortcomings of ICAP markets which include the bipolar nature of prices, minimal deliverability requirements and the lack of intrinsic value to customers, which necessitates reliance on administered demand. Furthermore, capacity payments are self-perpetuating since they tend to suppress energy prices, thus requiring supplemental payments to generators to cover their fixed costs and motivate new investment.

Motivated by these considerations, an alternative policy proposal focuses on energy availability rather than on capacity availability through contractual instruments that transfer investment and price risk between consumers and producers. The key difference is that an LSE is required to provide evidence that it holds long-term options or forward contracts that enable it to call sufficient energy or interruptible load to cover its peak load and reserves where needed. One advantage of this approach is that it addresses the actual need for energy in peak load conditions, and moreover, at a "strike price" for called energy that reflects the long-term elasticity of supplies. The second advantage is that the contractual obligations imposed on the LSEs provide intrinsic value to consumers by reducing their exposure to spot price risk while the deliverability condition entailed by the contracts imposes financial liability on the supplier that reflects locational energy prices in case of failure to perform. Furthermore, such obligations replicate prudent risk management practices that the LSEs would have undertaken in an idealized energy-only market where unmitigated prices that reflect scarcity rents drive investment decisions.[4] Hence, the contractual obligation may naturally become moot as the market matures and voluntary risk management by LSEs through long-term contracting and demand response results in adequate generation capacity.

C. Congestion Revenue Rights (CRRs)

CRRs, which can take the form of point-to-point obligations or options (FTRs) or flowgate rights (FGRs), are essentially financial instruments for hedging against congestion charges. Nevertheless, engineering standards have important effects on their specification and prices in order to facilitate their use as property rights to existing transmission capacity. Prices are affected by the requirement that in the aggregate they must be jointly feasible[5] and in the case of point-to-point rights, whether a CRR entails the obligation to pay when the nodal price difference is negative. Joint feasibility is an engineering standard that incorporates security and nomogram constraints in addition to the established available transmission capacity (ATC) and reflects mainly ratings based on thermal limits. But so too is the payment obligation, since it reflects the fact that flows counter to the direction of congestion reduce the amount and cost of redispatch needed to eliminate congestion. Engineering considerations are also relevant in other ways. For example, the distinction between the transmission capacity auctioned annually and the residual capacity auctioned monthly reflects an engineering judgment about how much capacity is surely available over the ensuing year. Similarly, the power transfer distribution factors (PTDFs) used in the optimization reflect engineering predictions of the PTDFs that will actually materialize in real-time operations. By allocating CRRs based on expected PTDFs, the TSO is essentially providing buyers of CRRs with hedges against variations in the PTDFs in actual operation of the grid. In sum, even though CRRs are financial hedges rather than contracts for physical delivery, they are crucially affected by engineering standards and judgments. Inevitably, therefore, market participants are concerned that the TSO applies engineering considerations in a way that maximizes the commercial value of these financial instruments.

The payment to a point-to-point CRR is the difference between the nodal prices at the two points. In turn, the vector of nodal prices is the product of the matrix of PTDFs and the vector of shadow prices on individual transmission elements (lines, transformers, etc.) derived from the optimization of power flows. From an engineering perspective, it is these shadow prices (called flowgate prices) that are the basic measures of the scarcity values of the transmission assets. Recognizing this, FERC's SMD retains provisions for flowgate pricing, and for CRRs that pay these prices—and because flowgate prices are never negative, they do not entail obligations to pay.

A commercial argument for flowgate pricing and CRRs based on them is that flowgate CRRs can be traded in secondary markets, whereas the market for point-to-point CRRs is too thin to be competitive—which is why the TSO must administer a frequent "reconfiguration" auction. A more basic engineering viewpoint follows from the observation that typically only a few major lines or inter-ties are congested recurrently, so only their flowgate prices are positive. In such cases the dispersion of nodal prices reflects mainly the dispersion of PTDFs, since most of the flowgate prices are zero. For purposes of financial hedging, therefore, the chief objective is to hedge against the flowgate prices of those lines subject to recurrent congestion. This is possible when the auctions of CRRs are designed to allow bids for flowgate rights that

[4]The Australian market is an energy only market where spot prices can rise to $8000/MWh. On July 15, 2005 the Public Utility Commission of Texas has voted to follow the Australian example and adopt an energy-only market approach for its generation adequacy provision (PUCT Project no. 24 255—Rulemaking Concerning Planning Reserve Margin Requirements.).

[5]Joint feasibility means that if all the financial right were simultaneously exercised through scheduling of corresponding physical flows, they could be accommodated by the available transmission capacity.

match what the engineers know to be the actual scarce transmission assets. That is, injection and extraction at nodes are not the actual constraints in the system; rather it is transmission lines and other elements. The fact that nodal prices mask the location of the binding constraints in the grid diminishes their usefulness as signals to stimulate incremental and new investments in transmission upgrades.

D. Grid Planning

Restructuring has severely altered the long-term planning process for new and incremental investments in the transmission system. In part, new obstacles arise due to tensions between federal and state regulatory jurisdictions, but most importantly they stem from the absence of any single entity able to undertake integrated resource planning for investments in new transmission and generation assets. The basic problem is the lack of a coordination mechanism and the limited ability of market mechanisms to fill this void. Those investing in transmission cannot be sure about the location and magnitude of new generation, and equally, investors in new generation may be uncertain about the future topology and capacity of the transmission grid.

The TSO invariably plays an important role in transmission planning because regulators and utilities now rely on it to identify where expansion is needed or useful. Expansion can improve reliability but it also affects generators and LSEs differently, and the impact varies by locations. Consequently, the TSO is also the one best able to measure the benefits and costs to various parties as well as the benefits from better reliability.

The TSO can also play an important role implicitly as the coordinator of investments. This can be done by establishing a long-term plan for grid enhancements, indicating how the topology will develop and what the proposed capacities should be, both for reliability and for efficient generation to meet the pattern of predicted load growth. Such a plan can anticipate strategic generation investments in response to transmission expansion and recapture some of the lost gains from abandoning the integrated resource planning approach by being proactive and developing transmission investment plans that will lead rather than follow generation investment.

A long-term plan for transmission expansion is the natural coordination mechanism because the TSO has the engineering resources to develop authoritative plans, and because generation investments have shorter lead times than transmission investments (and fewer regulatory hurdles and siting difficulties), investors in generation are better able to adapt their strategies to an established transmission plan than *vice versa*.

E. Introduce Contingencies Into Optimal Scheduling Models

Before restructuring, some utilities used a dynamic programming model to optimize operations, often over a 168-h rolling horizon, and the New England power pool and its successor TSO did so too. Most TSOs now use a single comprehensive day-ahead optimization for each day, supplemented by a unit commitment process and an OPF in the real-time balancing market. But here we focus less on the time horizon of the optimization and more on the role of uncertainty.

The current practice is to allow for uncertainty by including security constraints (e.g., N-1 and nomogram constraints) to guard against cascading failures, and by ensuring the reserve capacity deemed adequate, according to industry standards set by NERC and the regional councils, to guard against load surges. In addition, based on years of prior experience accumulated by utilities and power pools, each TSO has developed protocols for dealing with various specific contingencies that might arise. Such protocols generally follow the engineering principles for responding to emergencies, restoring stability to the grid, or recovering from a collapse over a wide area. The general thrust of these reliability principles is to put in place sufficient resources to handle all credible contingencies, complemented by skilled operators trained to follow standard procedures.

This method of dealing with uncertainty is time-tested and reliable from an engineering perspective. But experience has shown that it can have unintended consequences in the TSO's markets. We already mentioned the problems in New England due to depressed energy and reserve prices. The source of such problems lies ultimately in the optimization software used for unit commitment and dispatch, which we now explain briefly.

All the TSOs use optimization software that relies on a single prediction about the sequence of events that will unfold over the next day. Account is taken of possible variations from that prediction only by scheduling the required amounts of reserves in the various categories. This being the case, it is inevitable that the optimization sees no value from attributes that enable flexibility, like the fast start and high ramp rate of a CT. It is true of course that the categories of operating reserves (regulation, spin, etc.) recognize these attributes approximately, and with additional categories the software might be adapted to show a preference for flexible units that reflects the engineers' valuation of flexibility.

The preferable approach, however, is to initiate development of new software that accomplishes directly what the engineers want. The simplest form of "adaptive" optimization is based on a two-stage model, which actually represents fairly accurately the real situation. The first stage models the unit commitment and scheduling decisions taken day-ahead, whereas the second stage models the adaptations chosen by the operating engineers in real-time (balancing, calling on reserves, etc.). The essential ingredient is that, whereas the first stage represents irreversible commitments based on what is known day-ahead, the second stage appears in multiple versions, each corresponding to a different scenario about the events that have occurred since the first stage—changes in the weather, facility outages, etc. The objective of the optimization is to minimize the sum of the costs incurred in the first stage and the expected costs of remedying the deficiencies in the first-stage schedules revealed by the scenario that occurs.

An adaptive optimization recognizes the value of flexible resources. For example, a CT that would not be scheduled

based on a single prediction about the next day may be scheduled by an adaptive optimization. This is so because the optimization recognizes that assigning a CT to reserve status brings the advantage that it is available to respond quickly to outages or load surges that are contingencies represented in some scenarios—not the most likely scenario perhaps, but in some scenarios in which a fast response is very valuable if and when it occurs.

An adaptive optimization requires a faster and bigger computer in order to include multiple scenarios, but steady improvements in computing speeds suggest that engineers should start now to develop software that can provide this capability. From the more general perspective of engineering principles, adaptive optimization is inherently better because it schedules units based on their capabilities and limitations in a wide range of likely scenarios. Relying on a single prediction about the future and then scheduling reserves to meet an industry standard for "security-constrained economic dispatch" has been largely successful in protecting the grid, but improvements in both reliability and cost are possible by tailoring schedules more precisely to the range of contingencies that might occur.

F. Develop Seamless Interfaces Among Adjacent Systems

The proliferation of energy trading over wide regions has brought "seams issues" to the fore. Tags and transmission load relief (TLR) procedures are becoming obsolete as the TSOs within entire interconnections become more closely coordinated. However, even though communication capabilities now enable tight coordination, a methodology has not been developed for accomplishing it systematically. Ideally, exchanges of information among TSOs would allow the optimization by one TSO to be compatible with the optimizations by its neighbors. Even if one allows a lesser standard, it is still the case that there are no well-developed methodologies for ensuring that loop flows are fully accounted for, and that neighboring systems assign the same nodal prices at inter-ties that connect them. From the perspective of optimization theory, coordination among adjacent TSOs is an example of decomposition; i.e., the goal is to decompose the overall optimization for the entire interconnection into subproblems solved by each TSO separately, but coordinated via exchanges of information (such as net flows and nodal prices at boundaries).

VI. Conclusion

We described the interface and persistent gap between market and system operations in restructured electricity markets along with some lessoned learned through the restructuring experience in the United States and abroad. We sketched a list of potential reforms to commercial grid management strategies that define a research agenda aimed at bridging that divide.

Acknowledgment

Opinions expressed in this paper are those of the authors and do not represent the position of EPRI.

References

[1] "The NERC Functional Model—Functions and Relationships for Interconnected Systems Operation and Planning," Jan. 20, 2002 [Online]. Available: www.nerc.com

[2] W. Hogan, "Contract networks for electric power transmission," *J. Regulat. Econ.,* pp. 211–242, Dec. 1992.

[3] H.-P. Chao, S. Peck, S. Oren, and R. Wilson, "Flow-based transmission rights and congestion management," *Elect. J.,* pp. 38–58, Oct. 2000.

[4] FERC Staff Report, Principles for efficient and reliable reactive power suply and consumption Docket no. AD05-1-00, February 4, 2005.

[5] P. Cramton and S. Stoft, "A capacity market that makes sense (working paper)," Mar. 22, 2005 [Online]. Available: http://www.cramton.umd.edu/papers2005-2009/cramton-stoft-a-capacity-market-that-makes-sense.pdf

[6] S. S. Oren, "Ensuring generation adequacy in competitive electricity markets," in *Electricity Deregulation: Choices and Challenges*, M. J. Griffin and S. L. Puller, Eds. Chicago, IL: Univ. Chicago Press, 2005, ch. 10.

[7] R. Billington and R. N. Allan, "Power system reliability in perspective," in *Applied Reliability Assessment in Electric Power Systems*. New York: IEEE Press, 1991, pp. 1–6.

[8] M. D. Ilic, J. R. Lacalle-Melero, F. Nishimura, and W. W. Scenler, "An engineering-based approach to short-term economic energy management in a deregulated environment," Lab. Electromagn. Electron. Syst. Tech. Rep., Jun. 1991.

[9] North American Electric Reliability Council, "Reliability assessment 1997–2006," Oct. 1997 [Online]. Available: www.nerc.com

[10] N. J. Balu, T. Bertram, A. Bose, V. Brandwajn, G. Cauley, D. Curtice, A. Fouad, L. Fink, M. G. Lauby, B. F. Wollenberg, and J. N. Wrubel, "On-line power system security analysis (invited paper)," *Proc. IEEE,* vol. 80, no. 2, pp. 262–280, Feb. 1992.

[11] B. Stott, O. Alsac, and A. J. Monticelli, "Security analysis and optimization," *Proc. IEEE,* vol. 75, no. 12, Dec. 1987.

[12] C. A. Canizares, H. Chen, and W. Rosehart, "Pricing system security in electricity markets," in *Proc. Bulk Power Systems Dynamics and Control-V* Onomichi, Japan, 2001, pp. 1–11.

Hung-Po Chao received the Ph.D. degree in operations research and economics from Stanford University, Stanford, CA, in 1978.

He is Senior Technical Leader of Power Delivery and Markets at Electric Power Research Institute (EPRI), Palo Alto, CA, and Consulting Professor at Stanford University, Stanford, CA. He is responsible for the Transformation of Global Electricity Market Initiative (TGEM) at EPRI and has been engaged in the investigation of market design and restructuring policy, advising power companies, governments, and regulators in the United States and around the world. He has published over 50 papers and ten books/reports in this area.

Shmuel S. Oren (Fellow, IEEE) received the Ph.D. degree in engineering economic systems from Stanford University, Stanford, CA, in 1972.

He is a Professor in the Industrial Engineering and Operations Research Department at the University of California, Berkeley. He is the Berkeley site director of PSERC. He has been a consultant to private and government organizations including the Brazilian Electricity Regulatory Agency (ANEEL), the Alberta Energy Utility Board, the Polish system operator (PSE), and the Public Utility Commission of Texas (PUCT).

Prof. Oren is a Fellow of INFORMS.

Alex Papalexopoulos (Fellow, IEEE) received the Ph.D. degree in electrical engineering from the Georgia Institute of Technology, Atlanta, in 1985.

He is President and Founder of ECCO International, San Francisco, CA, a company that provides consulting services on electricity market design, power systems, energy management systems, and software issues to a wide range of private and public organizations in the United States and abroad. He was director of the PG&E's Electric Industry Restructuring Group in San Francisco, California until 1998. He published many articles and has made substantial contributions in the areas of network grid optimization and pricing, market design, implementation of EMS applications, and real-time control functions.

Dr. Papalexopoulos is the 1992 recipient of PG&E's Wall of Fame Award.

Dejan J. Sobajic (Senior Member, IEEE) received the Ph.D. degree from Case Western Reserve University, Cleveland, OH.

He was Director of Grid Reliability and Power Markets at EPRI. He is now International Consultant and President of Grid Engineering LLC. His research interests are transmission investment planning, market and grid operations, design of modern control centers, computational intelligence, etc.

Robert Wilson received the D.B.A. degree from Harvard University, Cambridge, MA, in 1963. The University of Chicago and the Norwegian School of Economics and Business Administration have conferred honorary doctorates.

He is a Professor Emeritus at Stanford University, Stanford, CA. He has worked on auction and market design in infrastructure industries such as telecommunications, gas transmission, and electricity, including advisory roles at EPRI and several system operators.

Prof. Wilson is an Elected Member of the National Academy of Sciences.