"Markets for Reliability and Financial Options in Electricity: Theory to Support the Practice"

by

Tim Mount
Professor of Applied Economics
and Management
Cornell University
214 Warren Hall
Ithaca, NY 14853-7801
tdm2@cornell.edu

William Schulze
Professor of Applied
Economics and Management
Cornell University
301 Warren Hall
Ithaca, NY 14853-7801
wds3@cornell.edu

Richard E. Schuler
Professor of Economics and of
Civil and Environmental Engineering
Cornell University
422 Hollister Hall
Ithaca, NY 14853-3501
res1@cornell.edu

Abstract

The underlying structure of why and how consumers value reliability of electric service is explored, together with the technological options and cost characteristics for the provision of reliability and the conditions under which market mechanisms can be used to match these values and costs efficiently. This analysis shows that the level of reliability of electricity provided through a network is a public good within a neighborhood, and unless planned demand reductions by customers have the identical negative value as an unexpected service interruption, market mechanisms will not reveal the true value of reliability. A public agency must determine that value and enforce the reliability criteria. Furthermore, in order to get an efficient level of demand response by customers in periods of system stress, they must see real time energy prices plus they must be paid an amount equal to the suppliers' cost of adding reliability to the system, if that amount is not included in real time prices.

An illustration is provided of how VARs might be scheduled and priced in contributing to system reliability, and a co-optimization procedure is required to determine energy and reserves simultaneously, similar to the method proposed by Chen, Thorp, Thomas, and Mount [1] for locational reserves. The optimization can be decomposed into a two step process – first, both required capacity and energy are selected based upon suppliers' offers over both dimensions through the minimization of expected costs over the list of contingencies necessary to satisfy the reliability criteria. This first step commits the reserves, but energy supplies are allocated in real time based upon the previous offer prices but the actual realized state of the electric system. This procedure which satisfies physical realities has a natural parallel in financial markets that have a forward option market with a strike price, followed by real time market clearing.

Proceedings of the 36th Hawaii International Conference on System Sciences (HICSS'03)
0-7695-1874-5/03 $17.00 © 2002 IEEE

1. Introduction

Electricity supply is comprised of at least three broad attributes that are valued by customers. However, in most instances these characteristics are bundled together with the flow of energy and sold at a combined price per mWh. From a customer's perspective these constituent parts are: (1) quantity of energy, (2) upon demand, (3) at a specified, consistent quality (e.g. voltage and frequency stability, low harmonic content, VARs upon demand). When partially unbundled, the electricity supplier may vary unit prices based upon the proportions of energy (1) and peak demand (2), since that influences the relative size of investment, but regulatory bodies normally set standards for the quality of service as gauged by (3), with some suppliers charging for kVARh usage, and suppliers average the costs for providing fewer outages than allowed by regulators in base prices. Unplanned outages can be thought of as a failure to satisfy category (2) services, since when the power is out, customers cannot satisfy their desire to consume as much energy as they want (1) at the quality they expect (3).

So when thinking about the reliability of electric service, it is comprised of category (3) and some aspects of category (2) of the electricity supply attributes that are valued by customers. Note that from a supplier's perspective, their response to these demand attributes may be through a different set and combination of mechanisms. As examples, providing reserve margins (spare capacity) of both generation and electric lines will satisfy both the customer's desire to use as much electricity as they want at the "flick-of-a-switch" and to not have their service interrupted (particularly unexpectedly). Furthermore, the provision of VARs while causing machinery to rotate, also can reduce the probability of an unexpected outage, so there is not a direct one-to-one set of mappings from the three demand attributes and the efficient supply responses.

Nevertheless, for the purposes of this analysis that focuses on potential customer actions, energy purchases will consist of category (1) and the portion of (2) required to respond to the customers' normal pattern of usage; whereas reliability will encompass category (3) and that part of (2) involved in mitigating unplanned service interruptions. In that light, the efficient set of prices, and consequently the desired structure of markets, are explored that serve the customers' desires, while being supplied through the least cost combination of available technological responses. Further, the time lags associated both with different supply responses (e.g. different generator ramp rates) and with the ease of customer adjustment, are compared and matched with a view toward devising an efficient sequence of markets that might be aligned with traditional financial paradigms.

2. Why Services Have Been Bundled

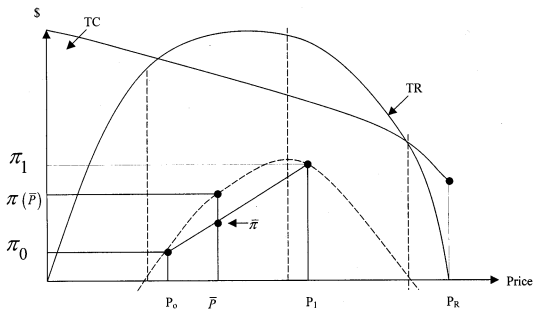
Allegedly, high costs of metering and avoidance of customer irritation are reasons given for not implementing widespread time-of-use (TOU) prices. Real time pricing (RTP) is simply an extension of the TOU concept over a much finer time-grain, and it is even rarer in practice. VARs are rarely billed separately, in part because up to a point they can be generated with no additional cost in conjunction with the supply of energy, but above a given proportion, further supply becomes quite expensive. And, customers rarely are asked what they are willing to pay for reliable service; reserve margins are provided according to regulatory criteria and the costs are averaged into the price of supplying energy. The exception is the occasional availability of "interruptible" rates or payments for "emergency demand response." But these pricing mechanisms are different than charging for reliability: the customer has some warning about when they will be asked to reduce load, in some instances the load reduction is still voluntary when requested, and only a small proportion of customers are involved in these programs in the limited areas where they are available. Why are prices not unbundled and variable widely over space and time?

Over space, homogeneous prices within a particular utility's service territory, for a given customer class, has been a tradition promoted by nearly every regulatory body, presumably under the theory that their legislative mandate that prices be "fair and non-discriminatory" means that they should be the same for all of a supplier's customers (never mind that the cost of service may vary widely, geographically). Here regulatory tradition, and a peculiar sense of what's fair, is a significant explanation.

Over time, why has RTP not caught on since the costs of generation vary so widely? One answer is to consider the likely effect on a supplier's (utility's) profits. Even with

decreasing costs, profit relationships as a function of price can be shown to be concave in prices as depicted in Figure 1, and if under real time prices P_1 and P_0 in peak versus low-load days, the supplier were to earn profits π_1 and π_0, respectively, then their average profit would be $\bar{\pi}$. If however, the supplier is allowed to charge the same average price in both periods, with a concave profit function, the firm would charge $\bar{P}$ and earn $\pi(\bar{P})$ which is greater than the result under real time pricing. This result is analogous to the Friedman and Savage [1] analysis of choice under uncertainty, except here the variability is assumed to be known. So one possible explanation for averaged, bundled prices is that it has been advantageous to suppliers.

Figure 1. Profits and Varying Average Prices



3. Theoretical Analysis of Optimal Supply of Reliability

Consider electricity customers whose value (utility for individuals; profits for businesses) is determined by the desired amount of electricity they consume, q^i, its reliability, $\bar{r}$, and the amount of all other goods (productive inputs for businesses), z^i, acquired. Also consider the possibility that customers can agree to reduce their use of electricity by an amount, Δ^i_r, upon request, in order to maintain system reliability, but that this agreed upon reduction in consumption also reduces the customer's value, so a more than offsetting compensation is required to induce this demand reduction behavior. By specifying the problem this way, the possibility that a decrease in reliability and an agreed upon demand reduction may not be the same can be explored, as well as circumstances when they may be similar since both do result in less use of electricity. What is different about the two is the fact that service interruptions are frequently unanticipated; whereas, advanced notice is given for demand reductions.

Furthermore, since there is a public good aspect to electricity reliability when service is provided through a network (even if one customer participates in demand reduction, but a neighbor doesn't, both will be interrupted identically if there is a system failure), the search for the optimal level of reliability must be modeled in the context of a centrally planned economy to allow for the optimal provision of services with public-good-like attributes, like reliability. In this context, the analysis demonstrates which decisions might be decentralized, allowing markets to provide efficient outcomes, and which require some public intervention (by the planner) to reach optimal outcomes. The planner's problem is to maximize society's welfare, which is enhanced by increased value to society's customers, subject to physical resource and technology constraints.

Formally, the planner's task is to solve the problem in equation (1). Each of the variables could be assigned a particular time designation so that q^i_0 might be i's desired electricity use in an off-peak period and q^i_1 might represent desired peak day usage, as an example.

$$
\begin{aligned}
&Max \;\; W\left\{U^i\left(q^i, \Delta^i_r, \bar{r}, z^i\right)\right\} \qquad (1)\\
&+\sum_j \lambda^j{}_1\left[z^j - c^j\left(q^j, r^j\right)\right]\\
&+\lambda_2\left[\sum_j g^j\left(r^j\right) + f\left(\sum_i \Delta^i_r\right) - \bar{r}\right]\\
&+\mu_1\left[\sum_j q^j - \bar{r}\sum_i q^i + \sum_i \Delta^i_r\right]\\
&+\mu_2\left[Z - \sum_i z^i - \sum_j z^j\right]\\
&0 \le \bar{r} \le 1
\end{aligned}
$$

Where: i = customers
j = firms
W = Society's welfare ranking
U = i's utility
q^i = electricity consumption
Δ^i_r = Agreed upon reduction times probability of selection
$\bar{r}$ = reliability
z^i = all other goods

In this specification, each supplier is thought to be able to produce greater supplies of electricity, q^j, and/or enhance it's reliability through r^j, either by providing reserves and/or VARS, as examples, but in so-doing the firm uses up resources (capital, labor and materials) so that less of the other goods can be produced, and z^j represents that offsetting reduction. Thus, each firm's cost of producing energy and reliability is represented by the reduction in z^j required to free the necessary resources, assuming efficient assignment of productive resources throughout the economy. In this way, each electricity supplier's use of resource is represented by its cost constraint (multiplied by Lagrangian λ_1^j) and society's aggregate resource constraint is represented by the maximum possible production of Z in the equation multiplied by Lagrangian μ_2, which is allocated to the consumption of z^i or the production of q^j and r^j through z^j. The other two constraints describe alternative means of producing electric service reliability, $\bar{r}$, in the equation multiplied by Lagrangian, λ_2, and in the equation multiplied by μ_1 which equates the desired demand for electricity, as adjusted by the realized reliability and agreed upon demand reductions, with the available supply. As specified, reliability is bounded by the unit interval and represents the probability the desired demand will be realized, if there is no call for demand reductions.

The first order conditions are summarized in equations (2) through (5) when the four constraints in (1) are binding. Here the λ^j multipliers are equal for all firms, so that as shown in equation (2), the production of energy should be allocated across all suppliers so that their marginal costs are equal.

$$\frac{\mu_1}{\mu_2} = \frac{\frac{1}{\bar{r}}\frac{\partial u^i}{\partial q^i}}{\frac{\partial u^i}{\partial z^i}} = \frac{\partial c^j}{\partial q^j} \qquad (2)$$

$$P(q) = MV(q^i) = MC(q^j)$$

$$\frac{\lambda_2}{\mu_2} = \frac{\frac{\partial c^j}{\partial r^j}}{\frac{\partial g}{\partial r^j}} \sim MC^i(\bar{r}) \quad \text{through } r^j \qquad (3)$$

$$\frac{-\frac{\partial u^i}{\partial \Delta_r^i}}{\frac{\partial u^i}{\partial z^i}} = \frac{\lambda_2}{\mu_2}\frac{\partial f}{\partial \Delta_r^i} + \frac{\mu_1}{\mu_2} \qquad (4)$$

$$\sum_i \frac{\frac{\partial u^i}{\partial \bar{r}}}{\frac{\partial u^i}{\partial z^i}} = \frac{\lambda_2}{\mu_2} + \frac{\mu_1}{\mu_2}\sum_i q^i \qquad (5)$$

where: MV(•) = marginal value
MC(•) = marginal cost
P(•) = price

Further, the marginal cost of supply should equal the customers' willingness to forego consumption of other goods in order to use more electric energy (their marginal valuation), which in turn should be equilibrated across all customers. Since the equalities in equation (1) are across all customers and suppliers, and in turn are equal to the ratio of the Lagrangian multipliers (analogous to the price of electric energy), this welfare optimizing solution's attributes show that similar results would be achieved by decentralizing the energy supply and consumption decisions by letting a market solve the problem. Note, however, if these decisions were indexed by time, and different marginal production costs and/or marginal values of use emerged at different times, then different prices are required for each time period.

Equation (3) requires that the service reliability provided by suppliers be allocated among them so that their marginal costs of production be equal. Equation (4) factors in the customers' contribution to service reliability and emphasizes that agreed upon demand reductions should be thought of as an alternative way of producing reliability, just as suppliers do so by providing greater reserves. To have an efficient level of demand reduction, equation (3) emphasizes that customers must receive a benefit equal to the sum of the foregone supply side alternative for providing that reliability, plus the

real time production cost saving for not having to supply that electricity. Only if they receive these combined savings will customers be induced to accept sufficient loss in the consumption value of electricity by dropping their energy usage to optimal levels. Optimal system reliability requires both demand reduction incentives and accurate real time pricing!

Equation (5) specifies the optimal level of reliability for the system, and the formulation emphasizes that the optimal level of reliability cannot be determined by a market; a regulatory mechanism is required to determine the sum of each customer's marginal valuations. If left solely to the market, each customer would equate their individual marginal benefits (not the sum) to the marginal cost, the price would be too low, and too little reliability would be provided. This is a classic public good problem, and while it is frequently difficult for a central authority to infer the sum of the valuations to customers of improved reliability, competitive markets will not get it right. Furthermore, even if suppliers were to provide reliability in order to further their own profitability, they would merely equate the two terms on the right-hand-side of equation (5), setting the marginal cost of providing additional reliability to its marginal benefit in terms of additional electricity sales. But equation (5) emphasizes that the cost of producing that additional electricity because reliability has increased is a cost, just as is the cost of supplying reserves, that must be added together and weighed against the sum of the marginal value over all effected customers.

Equation (6) compares the private valuation for demand reduction in equation (4) with the public demand for reliability in equation (5). While not the same, can the valuation for

$$-MV^i\left(\Delta_r^i\right)\bullet\frac{\partial\Delta_r^i}{\partial\bar{r}}=\frac{\lambda_2}{\mu_2}+\frac{\mu_1}{\mu_2}\bullet\frac{\partial\Delta_r^i}{\partial\bar{r}} \qquad (6a)$$

$$\uparrow \quad \uparrow$$

$$MC\left(\bar{r}\right) \quad P\left(q\right)$$

$$\downarrow \quad \downarrow$$

$$\sum_i MV^i\left(\bar{r}\right) \quad = \quad \frac{\lambda_2}{\mu_2}+\frac{\mu_1}{\mu_2}\bullet\sum_i q^i \qquad (6b)$$

demand reduction that is inferred by offering the proper incentives on the right hand side of equation (6a) be used to infer the private demand for reliability for use in equation (6b)? If it could, then these marginal values could be summed to infer the public demand for reliability. But note, even if the question in equation (7) is true, and it probably is not

Question: is

$$MV^i\left(\bar{r}\right)\sim\ -MV^i\left(\Delta_r^i\right)\bullet\frac{\partial\Delta_r^i}{\partial\bar{r}} \qquad ? \quad (7)$$

since planned demand reductions and unplanned interruptions may have quite different valuations, summing up the right hand side of equation (6a) leads to a very different cost than is represented by the right hand side of equation (6b). Thus the marginal value of reliability must be determined through surveys or the political process; market-derived information will not be adequate.

4. Example of Supplying Reliability: VARs

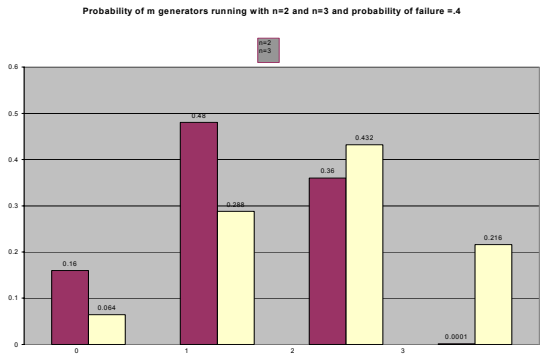
In establishing markets for electricity thus far, the conceptual problem of how to price VARs efficiently in terms of their contribution to improved system reliability has been ignored. In the context of equation (1), this contribution of VARs can be thought of as describing $g^j(r^j)$ in detail, and of understanding its cost characteristics in terms of improving system reliability, $\bar{r}$.

The setup for this example is an imaginary town that can import cheap hydro-power (q_h) from a considerable distance that costs c_h per mWh to produce. However, the town has a constant power factor so that it needs $X = \gamma Q$ mVAr to support load (where Q is total load) which must be supplied by local fossil fuel generators. Each of these identical fossil fuel generators has a cost of c_f per mWh for the power that each produces (q_f). We assume that the unit cost of fossil power is much more expensive than hydro-power, so $c_f > c_h$. Each of these plants has a minimum (q_f^{min}) and maximum (q_f^{max}) power setting. The reactive power produced by each generator (x_f) within this range is described as follows: $\alpha - \beta q_f \geq x_f$. Note that x_f is measured in absolute value units. If m fossil fuel generators are running out of n potentially available, then the total power available is $Q = mq_f + q_h$, and the total available reactive power is $mx_f \geq X$. Now consider the case where the probability of each of the fossil fuel generator failing is $(1 - \rho)$. If n generators are available, the probability that m will be running is then given by:

$$\phi(m,n) = \rho^m (1 - \rho)^{n-m} (m!)/(n!(n-m)!) \qquad (8)$$

Figure 2 compares the probabilities of m generators running when n=2 and n=3 for a value of $(1 - \rho) = .4$, admittedly a very large probability of failure, but useful for examining the properties of the distribution.

Figure 2.



If we treat $\phi(m,n)$ as continuous in the optimization problem to be defined below, the following approximations can be used:

$$\partial\phi/\partial n = ((m-n\rho)/(n-m))\phi, \text{ and} \qquad (9)$$

$$\partial\phi/\partial m = ((n\rho-m)/((1-\rho)m))\phi. \qquad (10)$$

From (9), the probability of having m generators running is increased by increasing n if $m > n\rho$ and decreased if $m < n\rho$. Note that $n\rho$ is the expected number of running generators. In Figure 2 above, with two initial generators, $n\rho = 2(.6) = 1.2$. Thus, adding a third generator decreases the probabilities of m=0 and m=1 since m<1.2 in these cases, and increases the probability of m=2 and adds a positive probability of m=3 since m>1.2.

From (10) the probability increases in m for $n\rho>m$ and the converse, as is approximately true for both the distribution for n=2 and the distribution for n=3 shown in Figure 2. Thus, (9) and (10) determine how system reliability ($\overline{r}$ through $g(r^j)$ in Section 3) is improved as the number of generators running, m, increases, and how in turn these probabilities change as the number of available generators, n, increase.

If we set this up as a static problem, the decision on n corresponds to how many fossil generators to construct to achieve an optimal level of reliability and the single period total cost for n plants is n times the capital recovery factor, CRF, times the investment cost, I. Thus, the total cost is

$$n(CRF)I.$$

The final component necessary to set up the optimization problem for choice of the number of fossil power plants to construct, n, their level of operation, q_f, and the amount of hydro power to import, q_h, is to define the benefits of power delivered as

$$B(Q) = \int_0^Q E(q)dq,$$

where B(Q) is similar to society's welfare function in equation (1), and E(q) is the inverse demand equation for power that determines the market price, E, as a function of the quantity of power produced. Note that, with downward sloping demand, $E' < 0$, and consequently, $B' > 0$ and $B'' < 0$.

To solve the optimization problem we start by maximizing the net benefits for each state of the world defined by m, the number of fossil generators in operation. Thus we wish to maximize

$$NB(m) = B(mq_f + q_h) - mc_fq_f - c_hq_h \qquad (11)$$

subject to the constraint on reactive power,

$$m(\alpha - \beta q_f) - \gamma(mq_f + q_h) \geq 0 \qquad (12)$$

(with the associated Lagrange multiplier λ), the constraint on minimum power setting,

$$q_f - q_f^{min} \geq 0 \qquad (13)$$

(with the Lagrange multiplier ψ), the constraint on maximum power setting,

$$q_f^{max} - q_f \geq 0 \qquad (14)$$

(with the Lagrange multiplier θ), and the constraint on hydro-power availability,

$$q_h^{max} - q_h \geq 0 \qquad (15)$$

(with the Lagrange multiplier ω).

For brevity we will discuss only the most interesting cases. First consider the case wherein only constraints (12) and (13) are binding. Note that as m varies, different constraints will become binding. This case arises when cost-minimization forces the minimum power setting

on fossil fuel generators to supply needed reactive power and hydro-power can meet demand net of m q_f^{min}, which is necessarily produced at higher cost to supply reactive power. In this situation, the first order conditions for q_h and q_f imply that (where we note that E = B′ which is similar to the market price)

$$E = c_h + \lambda\gamma, \text{ and}$$

$$E + \psi = c_f + \lambda\beta + \lambda\gamma.$$

Thus, the economic interpretation of these conditions is that hydro power is priced at cost plus a charge λ for reactive power, and fossil power operates at a vertex solution (at the minimum power setting) for each generator so $\psi>0$ and cheaper hydro power sets the market price. Here, if VARs are required only for providing short-run reliability on the wired network, the price for that reliability should be $\lambda\gamma$, and if only hydro firms are selling energy, they should pay λ for the reactive power that the energy they sell requires. Note that the term $\lambda\beta$ in the fossil generator optimizing condition represents the opportunity cost of producing less reactive power in order to produce greater mWhs.

Now consider the case where hydro-power is not able to meet demand so that (15) holds with equality. In this situation, ω is positive and fossil generators will optimally be run above the minimum power setting. The minimum power constraint no longer holds so ψ is equal to zero. Now the first order conditions imply that

$$E = c_h + \lambda\gamma + \omega, \text{ and}$$

$$E = c_f + \lambda\beta + \lambda\gamma,$$

so fossil generation sets the market price of electricity and ω can be interpreted as the excess unit profit earned by hydro given that more expensive fossil generation now sets the market price. Again, both hydro and fossil generation pay for needed reactive power that is incorporated into the bundled price of electric power paid by customers, and the price of reliability is $\lambda(\gamma+\beta)$.

The market structure implied by these conditions in this example where VARs are required as a constant proportion of energy suggests that fossil generators can sell both energy at a price of E - $\lambda\gamma$ and reactive power at a price of λ. Thus, the market price received for power by generators is adjusted downward to allow purchase of needed reactive power. This charge provides the revenues necessary to purchase reactive power from fossil generators. To provide appropriate incentives in each state m, the profit of the hydro generator should be

$$(E - \lambda\gamma)q_h - c_h\, q_h,$$

and the profit for each fossil generator should be

$$(E - \lambda\gamma)q_f + \lambda(\alpha - \beta q_f) - q_f\, c_f$$

where both E and λ are state dependant.

In summary, the optimal state dependant market structure requires that all generators be paid (E-$\lambda\gamma$) for each unit of power they produce and that fossil generators be paid λ for each unit of reactive power that they produce. Thus, we require two simultaneous related markets, one for power and another for reactive power.

The final issue is the determination of the optimal number of generators, n, for the system, that is the capital investment required to provide adequate reserves. Assuming constant static load, and that the market structure follows that outlined above so that the optimal level of net benefits can be obtained for each state of the world, NB(m)*, we wish to choose n to maximize

$$\int_0^n \phi(m,n)NB(m)^*dm \; - n(CRF)I.$$

Where NB(m) is the welfare function in equation (1), and reliability is improved by having more generators available, $\bar{r}(g(m))$. Thus, n is optimally chosen to satisfy

$$\varphi(n,n)NB(n)^* + \int_0^n (\partial\varphi/\partial n)\,NB(m)^*dm = (CRF)(I) \quad (16)$$

and an additional n^{th} generator should be added to the system as long as the sum of the expected net benefits with n generators plus the sum of the change in the expected benefits over all possible states of the system with respect to generator failure exceeds the annual capital cost of the investment. Note that, in general, adding a generator will reduce the probability of states where only a few generators are running and

increase the probability of states where more generators are running.

Note, that the optimizing condition for n includes net benefits. These incorporate the consumer surplus (welfare benefits) that will not be captured in the profits of individual generators using the market structure specified in this section. Either a governmental agency must make this optimizing decision about the desired level of reliability, n, or individual suppliers must be able to capture this consumer surplus. Thus, optimal private decisions on investment in generation implies that electricity must be sold by a discriminating monopolist who is free to capture the surplus, a difficult situation politically.

5. The Dynamic Process of Supplying Reliability

Although Section 4 lays out a mechanism to determine an optimal number of generators that should be available to maximize society's net benefits (welfare) by providing adequate VARs, since VARs are assumed to be required in fixed proportion to the energy required (mWh), these implied generation reserves are also available to meet load under a range of possible failure conditions. What is of interest is that the analytic development in Section 4 follows actual operational procedures in reverse. The analytics of Section 4 have us select the optimal level of VARs, given each state of the world (number of generators available, m), and then computes the optimal number of generators to make available, n, from the expectation over all states of the world, m. Of course operationally, the sequence is reversed and through a planning process (years ahead), the optimal number of generators are built, n; then a week or day ahead, adequate capacity is contracted for, turned on and ramped-up (again like narrowing a selection of n); and then in real time when the state of the world is revealed, an optimal dispatch is made over the available mix of generation, m. If the planning has been accurate, that generation should be adequate to meet load, despite contingencies (failures) with a probability of failing in only one day in one thousand (the current reliability standard specified by NERC).

Furthermore, the process described here is similar in sequence to the one implicit in the paper by Chen, Thorp, Thomas and Mount, 2003[2]. There, in solving for optimal locational reserves, the first step is to define the maximum and minimum calls for generation from each unit over the range of contingencies (failures and unanticipated demand), that must be satisfied in order to meet pre-specified reliability standards. Here again, the implicit assumption is that an over-arching authority specifies the reliability standard, $\bar{r}$ (e.g. solves equation (5)). Subject to offered prices from each generator for energy and reserves, the optimization criteria is to minimize the expected cost of meeting loads subject to the requirement that all load must be served under the specified contingencies and the maximum quantities of reserves and energy offered by each supplier. This optimization sets the reserve commitments and market clearing prices at each node, as well as an expected energy price at each location. However, in real time the state of the world (contingency, if any) will be revealed, and so in real time, it is optimal to select the energy dispatch and prices under the contingency that has been realized.

So either under the VAR analysis for providing reliability in Section 4, or in arranging for optimal locational reserves as in Chen, et. al [2], a multi-stage market is suggested where in the first stage the optimal amount of generation capacity is determined that guarantees meeting the reliability criteria at the lowest expected cost. Commitments are made for availability in this first stage and the prices for availability are set. Then in real time, given the state of the world, energy is dispatched at least cost and market prices for energy are set, based upon offers for energy in the first stage! This is an essential part of co-optimizaiton: suppliers must submit their maximum availability and their offer prices for supplying that availability and energy simultaneously. In this way competitive pressures are increased on suppliers to provide offers consistent with cost for both availability and energy, since their selection for either reserves and/or energy supply will be based upon both offers!

In an earlier analysis that focuses solely on the provision of reserves (e.g. 10 minute spin), Chao and Wilson [3] conclude that a two step clearing process is required to derive incentive compatible offers wherein actual energy prices are determined in the real time spot market and reserve prices and selection are set a period ahead. Because Chao and Wilson [3] presume that many other non-reserve-offering generators also offer into the real-time energy market (e.g. through day-ahead energy offers)

and the energy price offers by potential reserve providers are added to the other offers in selecting who will actually be called upon to generate in real time, the real-time energy market is assumed to be competitive. Given that condition, Chao and Wilson [3] demonstrate that it is incentive compatible to select generators for ten minute reserve availability solely on the basis of their reserve capacity price offers, since whether or not they will actually be selected to run in real time hinges on how their energy offers stack up with those of other suppliers in real time. This process differs from that of Chen, et. al. [2], who seek to derive a simultaneous least cost dispatch of both reserves and all energy through locational co-optimization. Achieving the potential efficiencies of locationally-specific reserve assignments requires that all generation with adequate ramping capability that indicates availability for energy supply also be available for reserve selection.

Under a co-optimization selection procedure, it is immaterial if the generator is asked for offers for reserves and for energy or for total availability and for energy, since the sum of reserves and generation quantities cannot exceed the total unit availability. If a two-stage sequence of auctions emerges, what is committed in the first stage (and what is required physically given ramp rates and required lead times to configure the system) is the total availability of each unit. Then in real time, it is efficient to minimize the cost of delivering energy, conditional on the energy prices offered and the capacity committed in the first stage. So the suggested sequencing is consistent with Chao and Wilson [3]; what differs is the basis for and nature of commitments made a period ahead under co-optimization. Furthermore, under co-optimization there is no assumption that energy markets will be competitive. In fact, a major objective of market experiments to be conducted on this assignment procedure is to assess the competitiveness of two part offers.

6. Concluding Observations

The preceding analysis suggests that the efficient provision of electric energy and service reliability could require a sequence of markets over time where the spacing is dictated both by the improvements in accuracy of information about actual demand (weather) and supply (equipment failure) conditions as the actual instantaneous transaction time approaches. All other commodity markets offer the opportunity for physical hedges against these uncertainties through storage and inventories, but that option is not widely available, physically for electricity. And so, the sequence of market information could be timed to match the sequence and lead time required for physical responses by suppliers, customers and the electric grid operators. Important considerations are the ramping rates of different types of generation available, the costs customers bear in rapidly adjusting to different market conditions as affected by the amount of advanced warning they receive, and the time required for the system operator to update their dispatch plan.

Starting with the real time market when the state of the world and all contingencies are known, if preceding markets have been properly designed, sufficient units should be ramped-up so that all that is required to balance supply and demand quantities is to make marginal calls for incremental energy supply and/or incremental demand reductions. The real time market is a pure energy market, with some latitude for variation provided by units on regulation.

Moving further back in time, in the preceding period all short term reserves will have been arranged in anticipation of contingencies, plus commitments for sufficient energy to meet anticipated load. Alternatively, availability payments could be made for the sum of each unit's reserves plus its expected energy sales, but the payment for energy sales could be deferred until the real time realization of energy actually required. Moving back in time, to insure adequate installed capacity, markets could again be conducted for availability, but once again those selections should be made based upon the minimization of expected costs of both availability and energy payments, where those expectations are computed over a range of contingencies and possible loads that together satisfy the system's reliability criteria.

What is proposed, therefore, is a sequence of co-optimizations, moving up to the real-time clearing. In each period except the real time, each supplier provides offer schedules with prices for both availability and energy where the availability is selected by a co-optimization calculation based upon both offer and demand schedules, where the availability quantity offered must equal or exceed the maximum quantity of

energy offered. What is paid in this period is the market clearing prices for availability. What is required of the successful offerer is that their total availability quantity offered in subsequent auctions must be the same or greater than in the previous ones, and their energy and availability price offers can be no higher than the successful clearing prices in the preceding auctions; otherwise the payments they received in the previous auction will be forfeited. Units not selected in earlier auctions would be free to revise their offer schedules without limit.

These markets would function, therefore, like a sequence of options markets where generators would be selected based upon the least cost combination of offer and strike prices – a two part pricing scheme. And to demonstrate the physical viability of earlier successful offers, once selected, suppliers would be required to offer into all subsequent markets quantities and prices no less favorable than received in earlier successful auctions, up until the final dispatch in real time. Thus a sequence of markets involving both capacity-availability and energy offers, whose selection is co-optimized in each period to provide the lowest expected combined cost of electricity, while serving the physical needs of suppliers, customers and operators of the system, would also have natural parallels in the financial community and could be represented as a sequence of binding options markets with maximum strike prices. There is ample room for further experimentation to explore whether real physical commitments must underlie each of these markets in a sequence, or whether pure financial exchanges are sufficient, up until the real time physical exchange of electricity.

References

1. Friedman, M. and Savage, L.J., "The Utility Analysis of Choices Involving Risk", Journal of Political Economy, 1948, p. 57-96.

2. Chen, J., Thorp, J., Thomas, R. and Mount, T., "Locational Pricing and Scheduling for an Integrated Energy-Reserve Market," paper presented, at Thirty-Sixth Hawaii International Conference on Systems Science, Jan. 6-9, 2003.

3. Chao, H. and Wilson, R., "Multi-Dimensional Procurement Auctions for Power Reserves: Robust Incentive-Compatible Scoring and Settlement Rules," EPRI and Stanford Univ., March 29, 2001.