MIT EL 99-002

Energy Laboratory

Massachusetts Institute
of Technology

Pricing Reliability: A Probabilistic Approach

May 1999

1999 Large Engineering Systems Conference on Power Engineering, June 20-22, Halifax, Canada (to appear).

Pricing Reliability: A Probabilistic Approach

José Fernando Prada
jfprada@alum.mit.edu

Marija D. Ilic
ilic@mit.edu

Massachusetts Institute of Technology, Cambridge MA 02139, USA

Abstract: This paper proposes the allocation of operating reserve in power systems through competitive capacity markets using a probabilistic approach. The insurance features of operating reserve are used to derive a valuation model that is analytically consistent and reflects the economic value of increased reliability to customers. The model can be expressed in the form of a demand curve for operating reserve. This curve can be used in auction-type capacity markets to determine the amount of reserve to be provided and its trading price.

Keywords: Operating reserve, spinning and non-spinning reserves, power outages, generation reliability, forced outage rate, capacity markets, willingness-to-pay, options, demand for reserves.

I. INTRODUCTION

The electric power industry has undergone major changes during the last decade and these changes have brought about new problems in power systems. Among them, the classical concept of power systems reliability needs to be reinterpreted in the context of the deregulated industry and of competitive power markets. In general, system adequacy and security are ensured by providing a series of ancillary services. The former problem of establishing reliability requirements in the regulated industry has turned into the question of what kind of mechanisms should be devised to allocate and price these services in competitive power markets.

This paper analyzes the allocation of the operating reserve at the system operator/power exchange level. The operating reserve is a fast-start capacity that must be kept available on a standby basis during normal operation to provide for unplanned outages of generating units.

System requirements for reserves have been traditionally defined using deterministic criteria such as "peak load percentage" or "loss of largest unit", which fail to consistently define the risk of supply shortages in the system. Furthermore, these and other quantity-constrained methods do not consider economic criteria and do not capture the worth of increased reliability provided by capacity reserves when they are employed in competitive markets.

We propose a market-based mechanism to allocate and price the operating reserve of the system using a probabilistic approach. The insurance-like features of operating reserve are used to derive a valuation model that is analytically consistent and reflects the economic value of increased reliability to customers. The model can be expressed in the form of a demand curve for operating reserve. This curve can be used in auction-type capacity markets to determine the amount of reserve to be provided and its trading price.

A. Background

Electric energy is produced and delivered practically on real time, without any convenient method to readily store it. For this reason, a continuous balance between production and consumption of electricity must be kept. Consequently, in power systems planning and operation, it is necessary to provide some generation margins above the expected peak load in order to cope with unexpected mismatches between supply and demand. Said margins are obtained by providing standby plant capacity.

Generation margins represent capacity reserves that can in turn be rapidly utilized in case of a generation shortage. Utilities have generally determined reserve requirements using working rules or in some cases probabilistic techniques. They estimate a reasonable amount of capacity to be reserved and kept standby, so that credible contingencies will not cause a failure of supply. Even when analytical methods are employed, a final decision regarding reserve levels depends on a judgment concerning the acceptable risk of failure. However, there are no simple rules governing this judgment. In essence, although not always made explicit, a decision concerning the tolerable risk of failure is a trade-off between the reliability offered and the cost of keeping reserves available. In regulated systems, where decision-making is centralized, the criterion used in practice is to offer high reliability provided that costs are not excessive.

B. Problem Formulation

The risk of shortages in power supply can be reduced by increasing the investment in generation and the operating cost of keeping reserves available. However, overinvestment and high operating costs would be ultimately reflected in the bill paid by the customer. On the other hand, underinvestment and tight generation margins would lead to a low reliability offered to customers. In general, economic efficiency requires that the benefits derived from improvements in reliability be weighed against the costs of providing additional reliability. Consequently, the main shortcoming of using quantity-based methods to estimate system reserve requirements, as previously described, is that economic criteria are not formally included in the decision-making process.

In theory, capacity markets can allocate system reserve efficiently. In such a market, the marginal benefit of increased reliability is made equal to the incremental cost of supplying capacity reserve. The market-clearing process defines both the amount of capacity to be reserved and the corresponding trading price. In a deregulated industry, therefore, a meaningful mechanism to allocate capacity reserve should be market-based. The market matches supply and demand, defines an efficient price for reserve and supports competition on the supply side, being consistent with the principles of economic deregulation.

Actual institutional set-ups and market designs offer practical approaches to capacity markets (e.g. UK and California). However, they use quantity-constrained methods to determine the reserve requirement and use this ex-ante figure as demand for reserves. These approaches provide little information about the value of reserves. Moreover, in markets where generators bid for supplying (making available) standby capacity, opportunities arise for strategic behavior of suppliers in the reserve market, and between the energy and reserve markets. In general, the main obstacle found in establishing markets for reserves is how to determine the value of the reliability benefits derived from additional capacity.

C. Proposed Approach

The valuation of capacity reserve is less straightforward than the valuation of energy. In effect, spare capacity is not a consumable good as is electric energy. Instead, what capacity reserve provides is a hedge against the contingency of not having enough generation available to meet demand. Essentially, a purchaser of capacity reserve holds the option to buy an amount of energy up to the quantity implicit in the 'locked' capacity, and he will do so according to the actual energy deficit confronted. A pricing method that did not consider these option- or insurance-like features would miss the real value of reserves. A suitable valuation model should associate the price paid for reserved capacity with the premium paid for holding the related option or insurance policy. Alternative mechanisms based on regulated rates or operating cost minimization would lead to less efficient outcomes.

This paper recognizes particular features found in reserves and proposes to use these characteristics to value reserves in capacity markets. The purpose of it is twofold. First, to create a suitable framework for operating reserve markets, and secondly, to introduce a pricing model to value reserves. The objective is to provide a more efficient approach to the allocation of operating reserve by taking into account its economic value. The gained insight will be helpful to regulators setting up market rules for capacity markets and to system operators (ISOs) or load aggregators who reserve capacity in behalf of electricity consumers. Both can use the proposed approach to make better-informed decisions in establishing market rules or purchasing reserves.

D. Paper Organization

The validity of a market-based framework for allocating operating reserve should be based on its consistency with technical requirements, with accepted criteria for economic efficiency, and on its feasibility of implementation in real systems. Accordingly, the paper is organized as follows:

- Section 2 reviews basic concepts of generation reliability and presents a mathematical model to evaluate reserve requirements.
- Section 3 discusses the trade-off reliability vs. cost, the benefits of reserve and the rationale for efficient allocation of reserve in capacity markets.
- Section 4 studies the value of operating reserve. We consider the benefits of having generation reserve in the system and we propose a method to evaluate those benefits. A pricing model is introduced to assess the worth of reserve in capacity markets.
- Section 5 presents conclusions and recommendations for further research.

II. GENERATION RESERVES AND RELIABILITY

Improvements in power systems reliability can be achieved by using better components or incorporating redundancy in the system. In generation, redundancy is obtained by providing spare capacity. Generation reserves are necessary to keep the risk of load demand exceeding available generation below an acceptable level. Generation reserves can be conceptually divided into installed capacity and operating capacity.

The installed capacity reserve relates to the long-term ability of the system to meet the expected demand requirements while the operating reserve relates to the short-term ability to meet a given load. Both must be considered at the planning level, but once an investment decision is made, the short-term requirement becomes an operating problem. The installed capacity considers the capacity that must be planned and constructed in advance to provide for (i) uncertainties in the forecast of demand growth, (ii) overhaul of generating equipment and plant maintenance, (iii) generation outages that are not planned or scheduled.

A. Operating Reserve

The basic difference between installed and operating capacity is in the time period considered. In the short term there is less uncertainty on forecast load. Moreover, equipment overhaul and maintenance can be scheduled during off-peak load periods. On the other hand, real-time balance of energy supply and demand, which is necessary to cope with load fluctuations, is achieved by frequency regulation or automatic generation control. Consequently, the operating reserve represents the capacity that must be available to replace loss of generation due to forced outages, which are events of stochastic nature.

Assuming there is sufficient installed capacity in the system, the allocation of operating reserve consists in the decision concerning which units to commit to replace failed generating units. In general, the risk of load interruption upon the failure of a generating unit can be minimized in two ways: (i) keeping part of the reserve 'spinning'; that is, as units connected to the grid, synchronized and ready to take load, and (ii) keeping available a group of units with quick-start capability. These units can be rapidly brought on-line and pick up load.

Both the spinning and non-spinning reserve form the operating reserve of the system. Non-spinning reserve can only be provided by hydraulic or gas turbine units which have start-up times in the order of minutes, whereas spinning reserve can be provided by a broader range of units. The division between spinning and non-spinning reserve can be somehow arbitrary and varies from system to system. In general, spinning reserve and non-spinning reserve available in less than 10 minutes are considered the fast-response contingent of the system, ready to replace units on outage.

B. Generation Reliability

The objective of generation reliability modeling is to derive suitable reliability indicators for the system on the basis of component failure data and system configuration. The indicators are essentially probabilistic estimates of the events leading to power supply shortages. The basic power system model used to evaluate the adequacy of a particular generation configuration is shown in Fig. 1, where appropriate generation and load models are combined to derive a risk model of supply shortages.

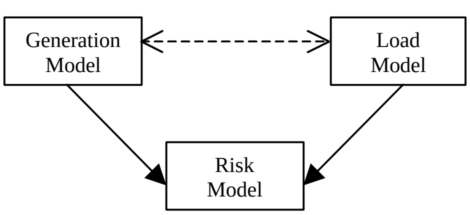

Fig. 1: Generation reliability evaluation.

Several approaches have been developed to carry out reliability studies, the most common are based on solving logical networks, fault tree analysis, solution of state-space models and Monte Carlo simulation. State-space representations are especially useful when evaluating generation reliability.

B.1. Generation Model

A model of generating units must consider the size of units and the two main processes involved in their behavior, namely the failure and the restoration process. A failure in a generating unit results in the unit being removed from service in order to be repaired or replaced, this event is known as a forced outage. Such outages can compromise the ability of the system to supply the demanded load and they have a significant impact on its reliability. To reflect this information, the status of generating units is generally represented by the states in which they reside.

This is a convenient representation because the state of each generating unit is continuously monitored and the duration of each state easily identified, so model data is available with enough accuracy. Thus, the operating life of a generating unit can be represented by a simple two-state model in a 'service-repair' process as shown in Fig. 2. The process consists of alternating 'up' and 'down' periods, T_U and T_D, both considered random variables. Perfect repair is assumed so the cycles are repeated during the useful lifetime of the unit. The operating history of the unit is determined by the probability distributions $f_U(t)$ and $f_D(t)$. Where $f_U(t)$ is the density function of up times T_U, that is, the durations of periods when the component is in service, and $f_D(t)$ is the density function of down times T_D, the durations of failed periods. If $\boldsymbol{X_t}$ is the state of the unit, up or down, at time t, then the following definitions apply:

- Probability of being up at t: $p_U(t) = p[\boldsymbol{X_t} = U]$
- Probability of being down at t: $p_D(t) = p[\boldsymbol{X_t} = D]$
- Mean up time, or mean time to failure (MTTF):

$$m = \int_0^\infty t.\,f_U(t)\,dt \qquad (1)$$

- Mean down time, or mean time to repair (MTTR):

$$r = \int_0^\infty t.\,f_D(t)\,dt \qquad (2)$$

- Failure rate: λ ; Repair rate: μ.

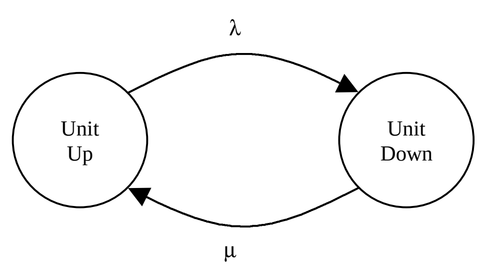

Fig. 2: Two-state model

- The mean time between failures (MTBF) is $T = m + r$

The most important quantity for generating reliability evaluation is the failure probability of the unit. For the simple two-state model the failure probability is given by the unavailability of the unit, U:

$$U = \frac{r}{m + r} \tag{3}$$

B.2. Two-State Model Analysis

The unavailability parameter can be expressed in terms of the unit failure and repair rates. Assuming both up and down times are exponentially distributed, with failure rate λ and repair rate μ, a solution for the state probabilities $p_U(t)$ and $p_D(t)$ can be obtained solving the two-state Markov process of Fig. 3. The corresponding probability density functions are given by:

$$f_U(t) = \lambda . e^{-\lambda t} \quad \text{and} \quad f_D(t) = \mu e^{-\mu t} \tag{4}$$

The probability of being in one particular state at t+Δt depends on the state the unit is at time t, not on the states previously assumed. Thus, the corresponding stochastic process is Markovian. The transition probabilities p_{ij} (from state i to j) are $p_{UD} = \lambda.\Delta t$ and $p_{DU} = \mu.\Delta t$. Consequently $p_{UU}=1-\lambda.\Delta t$ and $p_{DD}=1-\mu.\Delta t$, as indicated in Fig. 3.

- The state probabilities at $t+\Delta t$ are calculated as follows:

$$p_U(t+\Delta t) = p_U(t).(1-\lambda.\Delta t) + p_D(t).\mu.\Delta t \tag{5}$$

$$p_D(t+\Delta t) = p_U(t).\mu.\Delta t + p_D(t).(1-\lambda.\Delta t) \tag{6}$$

- Reorganizing (5) and (6), dividing by Δt and letting Δt → ∞ we obtain:

$$\mathbf{p}'(t) = \mathbf{p}(t) \bullet \mathbf{A} \tag{7}$$

Where $\mathbf{p}(t) = [p_U(t) \;\; p_D(t)]$ and $\mathbf{A} = \begin{bmatrix} -\lambda & \lambda \\ \mu & -\mu \end{bmatrix}$

- Solving (7), with the initial conditions $\mathbf{p}(0)$ = [1 0], we obtain the expressions (8) and (9) for $\mathbf{p}(t) = [p_U(t) \; p_D(t)]$ respectively.

$$p_U(t) = \frac{\mu}{\lambda + \mu} + \frac{\lambda}{\lambda + \mu} e^{-(\lambda + \mu)t} \tag{8}$$

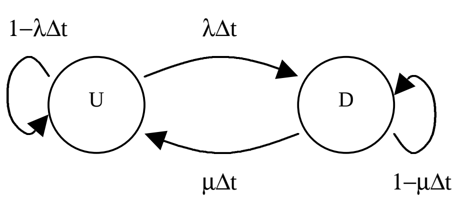

Fig 3: Two-state Markov process

$$p_D(t) = \frac{\lambda}{\lambda + \mu} - \frac{\lambda}{\lambda + \mu} e^{-(\lambda + \mu)t} \tag{9}$$

Using (9) we can write the unavailability of the unit as the long-run probability of finding the unit down, then:

$$U = \frac{\lambda}{\lambda + \mu} \tag{10}$$

B.3. Capacity Outage Distribution

The parameter U is a good approximation of the unit failure probability. The unavailability U is commonly referred to as the 'forced outage rate' (FOR) of the unit. The final step in building the generation model is to aggregate the generating units in the system to estimate available system generation. The available capacity of each generating unit is represented by a random variable with value 0 MW and probability u_i, and value equal to its nominal capacity g_i with probability $a_i = 1 - u_i$. For N generators present in the system, the available generation is:

$$\tilde{G}_A = \sum_{i=1}^{N} \tilde{g}_i \tag{10}$$

$\tilde{\mathbf{G}}_\mathbf{A}$ is a random variable itself describing the generating capacity available in the system. Assuming all units can fail and be repaired independently of failures and repairs of other units, the probability distribution of $\tilde{\mathbf{G}}_\mathbf{A}$ can be obtained combining the single probabilities of the individual units. The resulting distribution $\tilde{\mathbf{G}}_\mathbf{A} = \{G_j, p_j\}$, j = 1 … 2^N , represents a sample space of 2^N capacity states, where each capacity state represents an outage event with one or more units out of service. The capacity of a determined state j with k available units and N-k failed units is the sum of the capacities of the k available units, or $G_j = g_1 + \ldots + g_k$. The probability of the state is equal to the product of the probabilities a_i of the k units available and the probabilities u_i of the N-k units out of service, that is, $p_j = a_1 a_2 \ldots a_k . u_1 u_2 \ldots u_{N-k}$.

There are 2^N possible different capacity states. In practice, several states have the same capacity so they can be grouped in a single state with the same capacity and probability equal to the sum of the single probabilities. At the end the model is represented as a probability distribution with a series of capacity states and probabilities defined as follows:

$$\tilde{C}_A = \{C_j, p_j\}, \; \mathrm{j} = 1 \ldots l, \; l \leq 2^{\mathrm{N}}. \tag{11}$$

Where $C_j, p_j = \begin{cases} G_m, \sum p_m & \text{for } G_j = G_m \\ G_i, p_i & \text{for } G_j \neq G_i \end{cases}$

C. Generation Shortage Risk

The applicable capacity outage distribution should be combined with the system load to derive a measure of generation shortage risk. A loss of load will occur when the system load exceeds the generating capacity remaining in service. The simplest case is when the load is constant. If Lo is the expected load, the system loss-of-load probability will be the probability of all the outage events for which $C_A \leq L$o. The assumption of a constant load is sufficient for evaluating the adequacy of operating reserve, for example in systems where the dispatch is determined hourly. As indicated before, in the short term load uncertainty is small and load fluctuations are taken care of by load following services.

Another, more significant, measure evaluates the expected energy curtailed due to capacity outages. Effectively, from the point of view of the customers, the effects of outages resulting on an average 1 MWh curtailment would be very different from others resulting on an average 1,000 MWh curtailment, even if they had the same probability. We will keep the basic risk model expressed in the capacity outage distribution of (11) and use the incremental probability of energy curtailments to derive a valuation model for reserves.

III. RESERVES ALLOCATION

Due to the complex and integrated nature of power systems, failures in any part of the system can cause service disruptions. From the customer standpoint, power disruptions may be experienced as frequency and voltage reductions, unstable supply with erratic frequency variation, power fluctuations or the interruption of supply. Although all these events impose costs on customers, in practice the effects of supply interruptions are the most severe.

Ideally, supply should be made continuously available to customers, but that is costly and not physically feasible. In fact, interruptions of supply are caused by power outages, which are stochastic events involving the failure of one or several components in the system. Therefore, it is accepted that any system will present a definite risk of suffering a number of future power shortages. That is, unbalances between supply and demand that lead to load curtailments. The risk can be reduced by installing better equipment or by providing system redundancy. At the generation level, redundancy is provided as operating reserves, which can be dispatched in case of generation outages, effectively decreasing the probability of load curtailments.

A. Reliability vs. Cost

In order to reduce the frequency and duration of load curtailments, and lessen their effects on customers, it is necessary to incur the costs associated with keeping reserves available. As generation reliability is improved, a trade-off occurs between the increased costs of capacity reserves and the increased benefits to customers, as avoided interruption costs, from fewer power shortages. When making decisions concerning adequate levels of reserve, the factors to consider are, therefore, the incremental costs, the benefits expected and the allocation of capital and operating resources among the different parts of the system. The objective is to determine an optimal balance between the economic benefits accrued from higher reliability and the costs incurred by providing it.

In line with the trade-off reliability versus cost, there is a number of questions regarding the provision of capacity reserve: How much should be spent? What the optimal level of reliability is? Who should decide -power producers, regulators or customers-? On what basis should the decision be made? The answers to these questions represent the policy a system follows with regard to its generation reserves. The issue faced is not whether reserve should be provided at all, but rather by whom and how much. In the centralized decision-making of regulated systems, utilities have traditionally set reliability standards based on past operating statistics, without much emphasis on economic criteria. The result of this policy has been high levels of reliability offered, but uncertainty about the efficiency on the allocation of resources to generation reserves.

On the other hand, the reliable provision of generating capacity in a deregulated system depends essentially on the type of structure adopted. The ways in which it is accomplished will depend on the institutions devised to manage the system, the competitive mechanisms selected, the degree of coordination among system participants and the extent to which decision-making is decentralized. It is important to recognize that the conceptual base driving deregulation of the power industry favors competitive markets and decentralized decision-making as preferred mechanisms to allocate physical resources and foster individual choice. Furthermore, the trade-off between reliability and cost should be made explicit in order to determine a balance that is economically efficient.

B. Decision Criteria

The traditional criterion in systems where decision making is completely centralized has been to use least-cost resources in order to meet arbitrary levels of generation reliability. This sort of cost-effectiveness criterion implies an a priori selection of reserve levels, usually based on experience and judgment. Gains realized from different reliability levels are not considered.

A better approach compares the incremental cost of reserves with the corresponding decline in outage costs, that is, the economic costs incurred by consumers because of supply interruptions. This method minimizes investment and operating costs plus outage costs, over the period considered. The minimum cost allocation marks the optimal level of reliability to be used as a benchmark in the system. In this approach, the level of reserve is treated as a variable and total social costs, of both reserves supply and outage costs, are minimized. This is equivalent to a cost-benefit analysis that maximizes net social benefits. At the optimal level the incremental supply cost of reserves is equal to the marginal increase in benefits from avoided outage costs.

Despite difficulties on its application, the cost-benefit analysis is a valid economic approach, but it requires centralized decision-making. The cost-benefit approach does not incorporate individual choice in either supply or demand for reserves, being hardly compatible with a competitive electricity market, where suppliers prefer to decide individually the amount of capacity to commit. On the other hand, economics states that under certain conditions competitive markets lead to efficient outcomes. Competitive markets 'alone' allocate resources efficiently, without need of centralized direction. The pricing mechanism transmits the relevant information among market participants, allowing individuals to decide what is best for themselves. Market allocation is then economically efficient, it allows decentralized decision making and fosters individual choice. Therefore, a meaningful mechanism to allocate capacity reserve in deregulated systems should be market-based.

C. Market-Based Allocation

A competitive capacity market can allocate reserves efficiently. In such a market the marginal benefit from increased reliability is made equal to the marginal cost of supplying capacity reserves, thus maximizing net social benefit. When the market clears, it determines both the amount of capacity to be reserved R^* and the trading price for reserves. R^* defines the adequate level of reliability in the system, which is the one maximizing net benefit. A capacity market supports competition among reserve providers and sets the efficient price for reserves equal to the marginal cost of supply.

The market mechanism is shown in Fig. 4. The supply curve S represents the price at which suppliers are willing to make reserves available, and is equal to their marginal costs in competitive markets. The demand curve D_R indicates how much consumers are willing to pay for reserves, and is equal to the marginal value of reserves to consumers. At equilibrium, supply equals demand and the market settles at the clearing price P^* and the efficient reserves level R^*

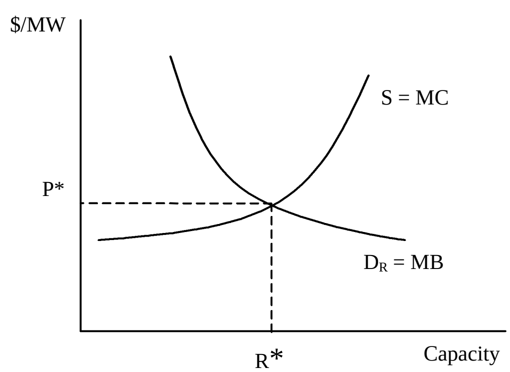

Fig. 4: Reserves market

C.1. Proposed Market Framework

The proposed market structure follows closely auction-type electricity spot markets, where hourly supply and demand bids for energy are submitted day ahead. The Power Exchange or System Operator (PX/ISO) collects the bids and clears the market by establishing the price and traded amount of energy hour by hour for the next day. In the reserves market, the PX/ISO will collect separate supply bids of capacity and will procure the service on behalf of the consumers. Afterwards, a schedule of operating reserve and prices for the 24 hours of the next day is produced. The PX/ISO intervenes only in its role of market maker and allocates the cost of purchased reserve among customers. This can be done on a simple pro-quota basis or any other appropriate method.

In the proposed framework the energy market is cleared first and the results represent the starting point of the reserves market. Competitive auctions are conducted for energy and operating reserve services. In the day-ahead auctions, individual generators are allowed to bid different hourly prices for energy and capacity reserves. Suppliers' energy and reserves bids are collected simultaneously, and they may offer the same capacity in all markets. Units with the lowest bids are selected to balance supply and demand in the different markets. The energy market is cleared first, defining energy output, price and dispatched units. The capacity committed in generation and regulation services is withdrawn from the operating reserve supply and the reserves market is cleared next. The result is a set of market-clearing prices and quantities for energy and reserves. Units left to supply reserve are not dispatched in the primary energy market so no energy opportunity cost is incurred.

C.2 Supply and Demand

The bidding to supply operating reserve is a transparent process. The tenders will contain prices and amounts of capacity at which spinning and non-spinning reserve would be made available during the next 24 hours. The bids of reserve will contain only a component for capacity. In case a reserve unit is effectively utilized, the energy consumed will be paid at the relevant spot price for the period during which the unit is kept generating. In theory, the energy spot price should be set by the energy bid of the marginal reserve unit dispatched on the system. Energy and capacity bids should be submitted at the same time to avoid strategic gaming between those markets. Assuming competition among power producers and no market power exercise, the bids should reflect the marginal cost of making operating reserve available.

The PX/ISO represents the aggregate demand side in the reserves market and procures the service on behalf of customers. Individual choice is favored by allowing consumers to participate in the market as interruptible loads. Consumers can choose to self-provide reserves or simply do not use them and sell them back. Interruptible loads are equivalent to generating capacity from the point of view of reserves. Thus, in case of a power shortage some load is (voluntarily) curtailed at a certain price, reducing the magnitude of the shortage. Bids for interruptible load can be prepared indicating the hourly amount of capacity that can be interrupted at a tender price for the next day. Interruptible loads can compete on equal terms with generation reserves, with the beneficial effect of enlarging the competitive base of the supply side. In the next section we discuss how the PX/ISO can establish the value of reliability to consumers and build a demand curve for reserves to confront with suppliers' bids.

IV. THE VALUE OF RESERVES

In power systems operation, generation reserves are used to provide for plant outages that could result in power shortages. In general, when load demand exceeds available supply and no corrective action is taken, the system becomes unstable. A dynamic process is triggered that can lead to unpredictable tripping of generators and lines, and widespread blackouts. In order to protect the system and the customers against a cascading blackout, it is necessary to establish an ordered reduction of load.

The load shedding helps to limit outage costs by protecting the integrity of the system and by decreasing the number of customers who would suffer interruptions of supply. Consequently, when load demand begins to exceed available supply, including generators' overload limits and maximum transfers from interconnected systems, a series of measures must be taken to reduce load. First, voluntary decreases or interruptible contracts would be called and at the end a power shortage will result in some load being cut off the system. The cost to customers of curtailed load is related to the amount of unserved energy during the curtailment period.

An optimal procedure of load shedding should minimize total social cost. This would imply to curtail first those customers that stand to lose the less from supply interruptions. We will assume then that load shedding could take place in an orderly way, in which loss of load is apportioned according an economic criterion to minimize social cost, where customers bearing lower losses from interruptions of supply are curtailed first.

A. Willingness To Pay

Power shortages result in interruptions of supply, which in turn translate into curtailments of energy to some customers. The effect of having capacity reserves in the system is to reduce the frequency and severity of said curtailments to consumers of electricity. Accordingly, the reliability benefits to customers will appear as reduction in costs associated with interruptions of supply, or outage costs. The method proposed in this paper measures outage costs based on willingness to pay for the unserved energy.

Considering the aggregate demand for energy, outage costs are measured as the reduction in net social benefit due to load curtailments. Net social benefit from electricity consumption is defined as the difference between what consumers are willing to pay for electricity and what consumers actually pays for it. The net benefit from electricity consumption or consumer surplus can be easily calculated if the demand curve for energy is known. With reference to Fig. 5, consumer surplus is the area *CS* below the demand curve and above the price line, as illustrated in figure 5

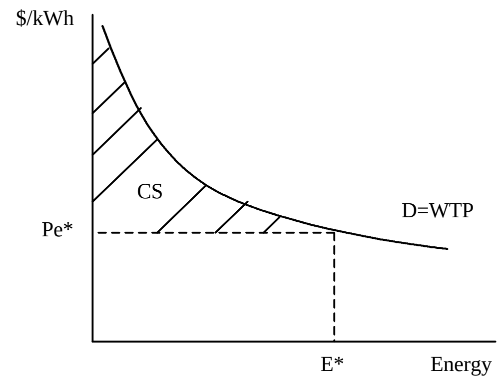

Fig. 5.: Consumer surplus

The demand curve D expresses the willingness to pay (WTP) of consumers for incremental units of electricity and measures the benefit derived form incremental units of energy. On the other hand, a similar concept or willingness to accept (WTA) expresses how much consumers would accept for giving up a unit of electricity. WTA measures the cost to consumers of decreasing units of energy. They should be compensated for this cost to accept reductions in consumption. Economic theory explains that the values of WTP and WTA should not differ. Consequently, the demand curve for energy represents the WTA for decreasing units of electricity, beginning from the market equilibrium quantity *E**, and can be used to measure consumer outage costs.

Load curtailments are assumed to be carried out according to a least-cost criterion. Therefore, they begin with the marginal consumption, indicated as the portion of the demand curve between points A and B in Fig. 6. The cost associated with curtailments is the lost surplus of marginal consumption, that is, the area under the demand curve net of the savings realized by not paying the energy price, or area ABC in Fig. 6. In conclusion, outage costs can be measured in terms of lost surplus of marginal consumers. The WTP approach is simple to apply and requires information readily available to the PX/ISO. In systems with demand side bidding the PX/ISO can directly calculate hourly outage costs to customers, due to unserved energy, by evaluating the expected surplus losses.

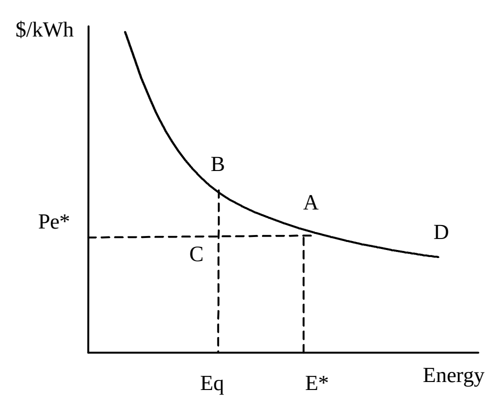

Fig. 6: Outage costs

In systems with no demand side bidding, a good approximation can be achieved using the market equilibrium information and an estimated price elasticity of demand, ε. For an isoelastic demand we can write the equation of the demand curve as $D = \alpha P^{\varepsilon}$, where $\varepsilon < 0$. Area ABC is given by the following expression:

$$\Delta ABC = \int_{Eq}^{E*}\left(\frac{\alpha^{1/|\varepsilon|}}{D^{1/|\varepsilon|}} - P_e^{*}\right)dD =$$

$$\frac{\alpha^{1/|\varepsilon|}}{1/|\varepsilon| - 1}\left[\frac{1}{E_q^{(1/|\varepsilon|-1)}} - \frac{1}{E^{*(1/|\varepsilon|-1)}}\right] - (E^{*} - E_q)P_e^{*} \quad (12)$$

Equation (12) is a good estimate for curtailments in the neighborhood of E^*, which is generally true because actual outages are expected to be small compared with total consumption. The main advantage of the lost surplus approach is that outage costs are directly derived from consumers' preferences as revealed in their demand for energy

B. Sources of Value

The value of capacity reserves is less intuitive than the value of energy. In effect, capacity is not a consumable good as is electrical energy, but the ability to produce this energy at a determined rate and with certain defined characteristics as cost, availability and rate of response. In fact, in power markets with advance scheduling, energy is actually traded through a series of forward contracts, while capacity payments correspond to option contracts on energy. Thus, in markets settled a day ahead for next-day hourly trades, the scheduled dispatch corresponds to a group of forward contracts to deliver energy a day later, at the specified market clearing forward prices. Likewise, payments made to available capacity represent premiums paid for energy option contracts, which can be called (dispatched) paying a specified amount.

B.1 Option value

Capacity contracts can be considered as call options on the energy (the underlying asset) that could be generated during the relevant dispatch period. The value of the underlying asset is by definition the spot price for energy. The strike price of the call option could be an associated energy price contracted or generator's own energy bid price. However, capacity reserves are called on because of generating outages and do not make part of hedging strategies against high prices. The result is that there is no well-defined exercise value as for 'financial' options.

Actually, the operating reserve can be considered as a strategic option. Thus, a purchaser of reserves holds the option to buy up to the amount of energy that can be generated using the 'locked' capacity. The decision to call the reserved capacity will depend on the actual energy deficit confronted. The value of reserved capacity to customers will be equivalent to the premium of holding the associated 'call' option. This premium essentially reflects the expected value of exercising the option, which is the expected value of curtailed energy associated with outage events.

B.2. Insurance Value

In theory, the risk of load curtailments could be insured, with customers being entitled to a financial compensation every time they suffer energy curtailments. The fair premium to pay for such an insurance policy would be equal to the expected economic loss faced. In practice, there are no insurance markets for power shortages. The other way to hedge against this risk is to procure reserves. In this case, instead of financial compensation, the physical risk of shortages can be reduced by increasing available capacity. We analyzed the stochastic nature of this risk in section 2. For this self-provided insurance, the customer will be willing to pay a premium (the fair price of risk) equivalent to the expected reduction in outage costs. This premium represents the worth of reserves for the customer, in contrast with the cost of acquiring them. The optimal level of reserve is determined when incremental reserve purchase costs are equal to the incremental savings in outage costs.

C. Reserves Valuation Model

In section 3 we proposed a practical market-based mechanism to define the adequate level of operating reserve in the system. In the proposed framework the PX/ISO acts as a single purchaser, procuring the service on behalf of customers. In order to establish the worth of reserves in the system, the PX/ISO must simply evaluate the expected reduction on outage costs due to additions of reserve units. This is equivalent to calculate the expected reduction in lost surplus of marginal consumption.

All the information required to evaluate the worth of reserves is a generation risk model as described in section 2 and the demand curve for energy, which is information known in systems with demand-side bidding. Otherwise, the curve can be inferred from the market clearing price and load, and an estimate of consumption price elasticity as in (12). This approximation should not introduce a big error for the range of interest, that is, for the size of expected capacity outages, which can be assumed to be small compared with the total dispatched generation.

Reserves Worth

Once the energy market is cleared the price of electricity P_E and the expected demand load L_D for the respective dispatch period, say 1 hour, are known. This information can be combined with the available capacity to establish the expected loss of consumer surplus during the hour. For the initial dispatch with no reserves, the capacity committed is equal to the expected load L_D. With reference to Fig. 7, for an outage resulting in a capacity level C_j, the corresponding power shortage is L_D - C_j with probability p_j. The unserved energy is (L_D - C_j) x 1h and the associated lost consumer surplus CS_j is the area below the demand curve D, between L_D and C_j, and above the price line P_E. The economic loss customers face is the consumer surplus CS_j. However, taking into account the uncertainty of the event, measured by p_j, the risk faced is the expected loss, p_j x CS_j. The total exposure for the aggregate customers is:

$$V_R = \sum_j p_j \, . CS_j \tag{13}$$

The worth of operating reserve is derived from reductions on this curtailment exposure provided by additions of fast-response capacity. The evaluation mechanism is as follows. The PX/ ISO, as purchaser of reserves, takes the bids, finds the expected available capacity for each unit using its forced outage rate, and builds the supply curve using the 'certain' equivalent megawatts. The expected available capacity of a unit, calculated as its availability times its nominal capacity, will be 'certain' in the sense of not being itself subject to outages. It will remove all curtailment risk for power shortages less or equal to the certain capacity. With reference to Fig. 7, the added value ΔVR_j of an amount of reserve R_j, covering shortages from C_{j-1} up to C_j, is simply the expected avoided loss p_j x ΔCS_j. The value of R_j is given by the sum of all the ΔVR_i for i < j.

$$VR_j = \sum_{i<j} p_j \, . \Delta CS_j \tag{14}$$

We illustrate the valuation method with the following example.

D. Example

Consider a power system consisting of six generating units with forced outage rates of 0.95. Unit 1 nominal capacity is 300 MW, units 2 and 3 are 200 MW each and units 4, 5 and 6 are 100 MW. System load is price-sensitive, described by the curve $D = 5000P^{-0.5}$. The market equilibrium is assumed to be at P_E = \$25/MWh and 1000 MWh. We want to know the value of reserves for this system.

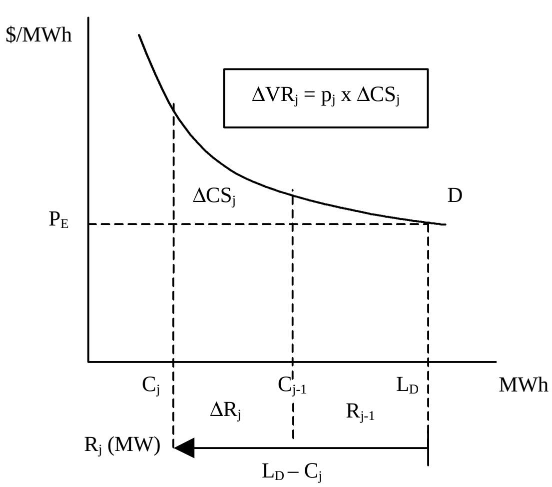

Fig. 7: Reserves added value

- The nominal capacity of the system is 1000 MW, the probability distribution of available capacity is shown in Table 1.
- Referring to Fig. 7 and using (12), the added value of reserves is given by:

$$\Delta VR_j = p_j \, . \int_{C_j}^{C_{j-1}} \left(\frac{5000^2}{D^2} - 25 \right) dD =$$

$$p_j \, . \left\{ 5000^2 \left[\frac{1}{Cj} - \frac{1}{C_{j-1}} \right] - 25.(C_j - C_{j-1}) \right\} \tag{14}$$

- The calculations for different levels of reserve are shown in Table 1.

Table 1 – Reserve Evaluation

C_J (MW)	PROB. P_J [$C_A = C_J$]	Δ SURP. (\$)	ADDED VALUE (\$)	RES. (MW)	RESER. VALUE (\$)
1000	0.73509	0	0	0	0
900	0.11607	277.8	32.2	100	32.2
800	0.08349	972.2	81.2	200	113.4
700	0.05101	1,964.3	100.2	300	213.6
600	0.00879	3,452.4	30.3	400	243.9
500	0.00473	5,833.3	27.6	500	271.5
400	0.00067	10,000.0	6.7	600	278.2
300	0.00014	18,333.3	2.6	700	280.8
200	0.00002	39,166.7	0.8	800	281.6

- The curve of operating reserve value in this system is shown in Fig. 8. As expected, the value of reserves rises rapidly at the beginning and after certain point there are diminishing benefits from additions of reserves.

E. Demand for Reserves

To establish how much operating capacity to reserve and the trading price for it, the PX/ISO needs to know the demand of customers for fast-response reserve. This demand is then confronted with the reserve supply curve derived from the capacity bids collected. The point where both curves intersect is the market equilibrium, which defines the optimal level of operating reserve in the system and its price. The demand curve measures the reliability benefits of reserve in the system and tells how much customers are willing to pay for reserves.

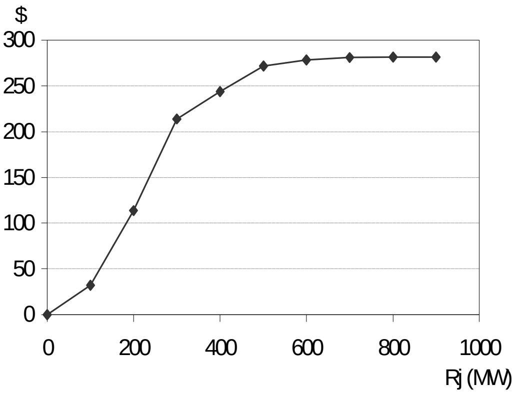

Fig. 8: Reserves worth

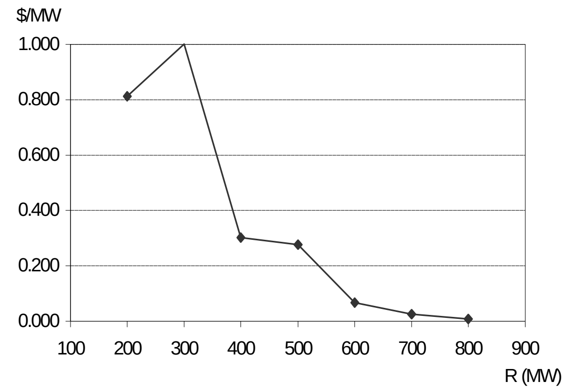

Fig. 9: Reserves demand

By definition, the demand represents marginal value consumers obtain from the consumption of a good. For reserves, therefore, it is the marginal reduction in expected loss for incremental units of available capacity. The demand curve shows then the incremental value of incremental units of reserve. For discrete increments of reserve we can determine the marginal value using the following relationship:

$$D_j = MVR_j = \mathrm{D}VR_j/\mathrm{D}R_j \;\text{ or }\; Dj = p_j.\mathrm{D}CS_j/\mathrm{D}R_j \qquad (15)$$

Using Table 1 and (15) we can derive a curve of reserves demand for the system of the previous example. The curve is shown in Fig. 9. The curve initially increases because the first MWs of reserve are highly valued. The increment in avoided lost surplus is the predominant factor here. Past certain point the curve becomes strictly decreasing, additional reserves have less value because the probability of using them becomes very small.

F. Markets for Reserves

A well-functioning capacity market for operating reserve is essential for the successful operation of modern power markets. It ensures not only the efficient provision of a basic reliability service but brings other beneficial effects by providing another source of cost recovery and by reducing strategic behavior among generators.

F.1 Cost Recovery

Marginal cost pricing ensures efficiency in allocation and the recovery of generation variable costs, but not necessarily the recovery of fixed costs. In theory, producer surplus in energy sales would yield enough revenue to recover fixed costs. However, base load units still may present some economies of scale. In addition, the intermittence of dispatch is also important. Peaking units that are dispatched few hours each year need high prices to recover fixed costs in a short period of time. Similar situation may exist regarding start-up costs. The stream of payments received for capacity reservation helps to recover fixed and start-up costs. For marginal units they represent a more predictable (although not more profitable) source of revenue than energy sales.

F.2 Strategic Behavior

When capacity is scarce, poorly designed market structures allow generators to speculate in energy and capacity markets, withdrawing capacity and manipulating bids to get higher prices. The proposed market framework reduces significantly those gaming opportunities. Having a single auction round for energy and capacity reserves and clearing the markets sequentially reduces he ability of generators to manipulate the bids. Moreover, the existence of a single energy bid price for the primary market and for dispatched reserves reduces arbitrage opportunities. A single energy bid makes sense because there is no difference whether a unit is dispatched in the primary market or called upon a contingency. The sole exception would be non-spinning reserve, where called units could be entitled to an extra payment to recover start-up costs.

F.3 Price Caps

Even in systems where market-based allocation of operating reserve is not favored, the concepts presented in this paper allow to establish ‘natural’, time-varying, price caps for reserve payments. With no markets for capacity, reserves could be created paying the full market price for energy without actually consuming that energy. Generators would be indifferent because they would receive at least its bid price with a high probability of saving fuel costs. Consequently, for arbitrage reasons, the maximum price that should be paid for reserves is the energy bid price of the marginal unit reserved. The rationale is that, in case the reserve were purchased as energy in the primary market, the marginal reserve unit would be the marginal unit in the dispatch, establishing a market clearing price equal to its bid. This price cap will adjust itself, hour by hour, to the conditions of the system.

V. CONCLUSIONS

This paper presents an alternative to the widespread practice of defining operating reserve requirements based on traditional deterministic criteria. This and other quantity-constrained methods rely on an arbitrary selection of the adequate level of reserves and do not address economic efficiency beyond some cost-minimization criterion. Moreover, when these methods are applied in competitive power markets, they could provide perverse incentives to power producers, which may explain anomalous reserve prices occasionally observed in deregulated systems.

The proposed market-based framework for allocation and pricing of operating reserves in power systems is economically efficient and fosters individual choice and competition in power markets. In addition, it reduces opportunities for strategic gaming and provides a simple mechanism to determine reserve levels and their trading price. We have introduced a valuation model, integral to the proposed framework, which allows a market maker or system operator to organize reserve markets in coordination with the energy market. The valuation model is based on the probabilistic assessment of the economic risk faced by customers due to power outages.

It is suggested that similar value-based methods could be developed for planning reserves of installed capacity and for other generator-provided ancillary services. Further research in these fields is encouraged.

VI. ACKNOWLEDGMENTS

The authors would like to acknowledge the financial support provided by COLCIENCIAS and the sponsors of the MIT Energy Lab Consortium "New Concepts and Software for the Electric Power Industry under Restructuring". The authors also appreciate discussions with Mr. Frank Graves and Dr. Richard Tabors.

VII. REFERENCES

[1] Prada, J. *Valuation of Operating Reserve in Power Systems*. MIT, M.Sc. Thesis, Technology & Policy and Electrical Engineering, June 1999.

[2] Endrenyi, J. 1978. *Reliability Modeling in Electric Power Systems*. John Wiley & Sons, New York, NY.

[3] Billinton, R. and Allan, R. 1996. *Reliability Evaluation of Power Systems*. Plenum Press, New York, NY.

[4] Hirst, E. and Kirby, B. "Ancillary Services: The Neglected Feature of Bulk Power Markets". *Electricity Journal*, Vol. 11, No. 3, Elsevier Science, April 1998, pp. 50-57.

[5] Hirst, E. and Kirby, B. "Technical and Market Issues for Operating Reserve". *Electricity Journal*, Vol. 12, No.2, Elsevier Science, March 1999, pp. 36-48.

[6] Billinton, R. 1970. *Power System Reliability Evaluation*. Gordon and Breach Science Publishers, New York, NY.

[7] Telson, M. 1973. *The Economics of Reliability for Electric Generation Systems*. MIT Energy Laboratory, Report No. MIT-EL 73-016, Cambridge, MA.

[8] Knight, U. 1972. *Power Systems Engineering and Mathematics*. Pergamon Press, Oxford.

[9] Munasinghe, M. 1979. *The Economics of Power System Reliability and Planning*. The World Bank, John Hopkins University Press, Baltimore, MD.

[10] Ilic, M. and Galiana, F. 1998. "Power Systems Operation: Old vs. New". *Power Systems Restructuring, Engineering and Economics*. Ed. by M. Ilic, F. Galiana, L. Fink. Kluwer Academic Publishers, Boston, MA, pp. 15-107.

[11] Kirsch, L. and Rajaraman, R. "Profiting from Operating Reserves". *Electricity Journal*, Vol. 11, No.2, Elsevier Science, March 1998, pp. 40-49.

[12] Hunt, S. and Shuttleworth, G. 1996. *Competition and Choice in Electricity*. John Wiley & Sons, Chichester, England.

[13] Borch, K. 1990. *Economics of Insurance*. North-Holland, Amsterdam.

[14] Graves, F., Grant, E. et al. 1998. "One-Part Markets for Electric Power: Ensuring the Benefits of Competition". *Power Systems Restructuring, Engineering and Economics*. Ed. by M. Ilic, F. Galiana, L. Fink. Kluwer Academic Publishers, Boston, MA, pp. 243-280.

[15] Hogan, W., Cadwalader, M. et al. 1998. *Reliability, Scheduling Markets, and Electricity Pricing*. Harvard Electricity Policy Group, Cambridge, MA.

[16] Tseng, C., Oren, S. et al. 1999. "Price-Based Adaptive Spinning Reserve Requirements in Power Systems Scheduling". *Electrical Power & Energy Systems*, Volume 21, Number 2, Elsevier Science, February 1999, pp. 137-145.

[17] Energy Information Administration. *Electricity Prices in a Competitive Environment*. DOE/EIA 0614, Washington D.C., August 1997.