DESIGN AND MANAGEMENT OF CURTAILABLE ELECTRICITY SERVICE TO REDUCE ANNUAL PEAKS

by

Shmuel S. Oren and Stephen A. Smith

Memorandum No. UCB/ERL/IGCT M90/65

26 July 1990

ELECTRONICS RESEARCH LABORATORY

College of Engineering
University of California, Berkeley
94720

Design and Management of Curtailable Electricity Service to Reduce Annual Peaks

Shmuel S. Oren
University of California Berkeley
Stephen A. Smith
Santa Clara University

Abstract

Curtailable electricity service is a voluntary option in which customers receive credits for permitting the utility a certain number of discretionary interruptions per year. This paper describes a methodology that allows an electric utility to design and manage these service offerings to achieve maximal peak load reduction. This methodology was developed as a case study project jointly sponsored by the Electric Power Research Institute (EPRI) and New England Electric Service (NEES). It provides a decision support tool for the system dispatcher, who must decide on a daily basis, when to call for an interruption; as well as methods for planning new service offerings, by determining the service attributes and pricing that will lead to the most efficient use of customer interruptions. Although the methodology was developed for NEES, it can be applied at other electric utilities, including those who use curtailable service offerings to reduce peak energy costs or to improve tight reserve margins.

Interruptible and curtailable service contracts for electricity have gained popularity over the last decade as load management techniques for reducing utilities' operating cost and mitigating capacity shortages. Many utilities offer discounted electricity rates to customers (usually industrial) who allow part of their load to be curtailed at the utility's discretion. The contracts typically specify the warning time, maximum duration and frequency of the curtailments. In some cases, customers can choose the terms of their contract from a menu that specifies the discounts associated with various available sets of options.

The models and analysis described in this paper were developed as part of a case study conducted at New England Electric System (NEES) which currently offers curtailable service called CIS (Cooperative Interruptible Service) to its large industrial and commercial customers. The case study was jointly sponsored by NEES and the Electric Power Research Institute (EPRI). Some related background details concerning NEES are outlined in Section 1. A more complete description and rationale for the program is contained in a working paper by Wharton (1988).

This paper considers two problems faced by an electric utility with curtailable service subscribers. (1) How should the existing contracts be used in the best possible way to reduce the utility's peak loads? (2) What is the incremental value of additional interruptible service contracts and what options should be offered? The specific objective function used in this paper is minimization of the utility's annual peak load. This was motivated by the needs of NEES that stem from the cost allocation rules of the New England Power Pool (NEPOOL) to which NEES belongs. Reducing the annual peak load is an important objective for all electric utilities, since capacity requirements and reserve margin requirements are largely dictated by peak loads. The same methods that we have developed in this paper can be applied to obtain the maximal increase in available reserve margin or to reduce the peak real time price or spot price that the utility experiences.

The economics literature dealing with design and pricing of interruptible/curtailable service contracts for electricity has addressed these issues from the point of view of product differentiation with respect to the attribute of service reliability. Some of the early work on differentiation and pricing of service reliability in the electric power industry and on the efficient rationing of power is due to Crew and Kleindorfer (1978), Fisinger (1980), Marchand (1974), Telson (1975), Tschirhart and Jen (1979) and Vickery (1971). More recent work has focused on the design and pricing of priority service and the alternative implementations which include interruptible and curtailable service contracts. This work is contained in the articles and reports by Chao et. al. (1986), (1987) (1989), Oren and Doucet(1990) Smith(1989) Viswanathan and Tse (1989).

The work presented in this paper differs from previous research in several respects. First, much of the earlier work on priority service and rationing is based on supply uncertainties, while here the focus is on the reduction of demand fluctuations. Second, while most of the previous work does not address the operational aspects of implementing a priority service program, a main focus of this paper is the operational dispatching of service curtailments. From this perspective, curtailable service contracts are a resource rather than a component of service reliability. Consequently, the management and design of the contracts becomes a problem of allocation and acquisition of a scarce resource to meet uncertain demand. A key objective of this research was to develop operational tools and decision aids for (1) the dispatcher, who decides what interruptions should be called for on a particular day, and (2) the rate design manager, who decides what menu of service options and prices to offer. An important consideration in this development has been the availability of information and the ease of use of the operational tools.

The first part of the paper describes a methodology used for dispatching the customer load reductions available from a given set of interruptible service contracts. The methodology is evaluated using simulation based on the historical load data from NEES. The second part of the paper explores the desirable mix of service contracts and their relative pricing so as to achieve a reduction in annual peak at the least social cost. Although the analysis is performed in the context of NEES's load patterns and service offerings, its results and insights should apply to other interruptible and curtailable service programs, which are currently offered in some form by most electric utilities. As noted previously, other electric utilities may replace annual peak demand with other metrics to be optimized over the year, such as peak real time price or reserve margin.

1. Background on NEES

New England Electric Service (NEES) offers a voluntary interruptible service program, known as Cooperative Interruptible Service (CIS) for its major industrial and commercial customers. CIS subscribers specify a firm power level or load, which must be at least 10% below their annual peak demand. Upon notification by NEES, subject to a specified warning time, each subscriber agrees to reduce his load below his firm power level for the next day. The period of load curtailment lasts throughout the normal business hours of the day, which constitute the peak period. Such calls to subscribers can be made at most 30 times per year under the current service offerings.

The primary purpose of CIS is to provide a means for reducing NEES's largest daily peak loads. Peak loads are a key factor in determining supply costs for all electric utilities. For the region served by NEPOOL, the New England Power Pool, which is the central supplier and dispatcher of power for NEES, peak loads also increase the risk power shortages, because generation capacity is tight in this region. NEES must pay a yearly demand charge to NEPOOL based on its annual peak load; thus peak load reduction results in a direct costs savings in this demand charge. Reduction in the largest daily peak demands also decreases the risk of power shortfalls, both for NEES's customers, and for NEPOOL as a whole. Because its warning time requirement is 12 to 16 hours, CIS cannot respond directly to unexpected supply shortfalls such as the failure of a generating unit, which must be handled by other procedures.

The pattern of daily peak demands at NEES is highly variable, as shown in Figure 1. There are two seasonal peaks: a Winter or heating peak that includes December, January and February; and a Summer or air conditioning peak, which includes June, July and August. There is virtually no chance of an annual peak in the other six months of the year. In addition to seasonal changes, weekend and holiday demands fall roughly 25% below the weekday demand in each season, which produces the periodic oscillations in the figure. Again, there is essentially no chance of an annual peak on a weekend or holiday. Within the peak seasonal months, there is considerable variation from one week to the next, which is largely due to the weather. A prolonged hot spell generally results in successively increasing demands for air conditioning as structures gradually heat up. Prolonged cold spells produce similar results in Winter. Thus, the accuracy of the daily peak demand forecasts is similar to the accuracy of the daily temperature forecasts for the service region.

NEES currently has roughly 100 MW of interruptible load subscribed to CIS, which, in principal, should allow its annual peak to be reduced by 100 MW. The decision of whether to call for an interruption is made sequentially each day, and it relies on three imperfect forecasts:

(1) The forecast of the annual peak
(2) The forecast of the daily peak for the next day
(3) The forecast of the number of future daily peaks that will be larger than the next day's forecasted peak.

Errors in any of these forecasts can cause interruptions to be allocated suboptimally. NEES was particularly interested in obtaining a dispatching methodology that could be characterized by a simple heuristic, which could be implemented in a spreadsheet program used by the system dispatcher.

2. Planning and Dispatching a Single Block of Interruptible Load

This section describes the solution for the case in which all the interruptible load is dispatched as one block, which is the current situation at NEES. We develop the planning methodology that addresses the question: For a given amount of interruptible load L, how many interruptions are needed for the year and what customer credits should be offered? Given that the total interruptible load is L, we need only be concerned with a band L below the annual peak, since that is the largest possible reduction in the annual peak. The planning model is designed so that the probability of missing a peak within L of the annual peak is no larger than some specified value α. We determine the number of interruptions k such that when they are dispatched optimally, the desired confidence level is achieved.

The dispatching model serves as a decision support tool for the system dispatcher who must decide a day in advance whether to call for an interruption. Given k remaining interruptions at any point in time, the best dispatching policy is to attempt to curtail the k highest remaining peaks. At any given time, the utility knows the seasonal distribution of daily peaks, the number of interruptions yet remaining for each class of customers, the number of peak demand days yet remaining, the estimated yearly peak and the daily peak forecast for the next day. The dispatching methodology calculates a threshold such that an interruption is called for if the daily forecast exceeds the threshold. The threshold is updated dynamically each day, as the input information changes.

The dispatching problem is, in principle, a stochastic dynamic programming problem, where the state space includes the number of remaining interruptions and the number of peak demand days remaining. Since at any point in time the forecasted future cost depends on the highest peak (after curtailment) observed so far, the state space must also include the highest observed peak. This value is affected by the entire dispatching history. The resulting large state space combined with the uncertainties involved have made the dynamic programming approach intractable. Instead, we use an "open loop" approximation with a rolling horizon. This determines a threshold H for the next day's interruption decision under the assumption that the same threshold will be used throughout the remainder of the contract period. The threshold is computed by balancing the tradeoffs between (1) the probability α of missing a peak due to forecast error; (2) the probability β of running out of interruptions prematurely. While lowering H will reduce the probability of missing an important peak, it increases the probability of exhausting the allotted interruptions prematurely.

The derivation of the threshold level and the relationship between k, α and β discussed above, requires specification of the two probability distributions

$$P\{\text{yearly peak} = x \mid \text{forecasted to be } \mu\} = \phi\left(\frac{x-\mu}{\sigma}\right)\frac{1}{\sigma} \tag{2.1}$$

$$P\{\text{daily peak forecast} = F \mid \text{actual peak value} = x\} = \phi\left(\frac{F-x}{\sigma_1}\right)\frac{1}{\sigma_1}, \tag{2.2}$$

where $\phi(z) = \dfrac{e^{-z^2/2}}{\sqrt{2\pi}}$.

The probability densities (2.1) and (2.2) are assumed to be Normal with 0 means and to be stationary, in that they are independent of the seasonal time in the year.

Historical data was available at NEES on one-day-ahead forecast errors for the daily peaks. Figure 2 shows a histogram of the daily forecast errors for 1988 for the top 64 daily peaks, representing the typical days on which key interruption decisions will be made. The Normal distribution (2.2) is convenient for analytical purposes, and based on the p-value of the Chi-sqaure statistic for the data in Figure 2, the Normal distribution hypothesis would not be rejected at the confidence level of 10% and the mean of the daily peak forecast errors is not significantly different from 0 for this data. The unbiased estimate of the standard deviation σ_1 is 179 MW.

Historical data was not available on the errors in forecasting annual peaks and, in any case, its sample size would be insufficient for statistical analysis. For convenience, the Normal distribution (2.1) was assumed for the annual peak forecast error. Based on discussions with the NEES dispatching personnel, σ the standard deviation of the annual peak forecast error was subjectively estimated to be 100 MW. The total forecast error, which is the sum of the forecast error and the error in estimating the annual peak, thus has a standard deviation given by

$$\sigma_2 = \sqrt{\sigma^2+\sigma_1^2}. \tag{2.3}$$

[This value was determined to be 205 Megawatts in the case of the NEES data.]

To capture seasonal fluctuations, a weight factor w_t was determined for each day t of the year, defined as

$$w_t = P\{\text{day } t \text{ is a peak demand day}\}. \tag{2.4}$$

with $$W_T = \sum_{t=T}^{365} w_t = E[\text{ number of peak demand days remaining at day } T]. \tag{2.5}$$

A "peak day" is classified as such by the system dispatcher. These days can viewed as the set of possible candidates for interruptions. The relative frequency of "peak demand days" was determined for each month, based on data from the three year period 1986–1988. Weekends and holidays were assigned a weight of zero. Days with nonzero weights are assigned equal w_t values within each month determined so that the sum of the w_t in each month equals the expected number of peak demand days in that month based on historical data. This provides us with an estimate of the probability in (2.4).

The probability distribution for loads on peak demand days relative to the annual peak is defined by

$F(D)$ = Expected fraction of peak demand days within D of the yearly peak.

The distribution $F(D)$ was calculated empirically, based on the data for peak demand days for the period 1986-88. The data, illustrated in Figure 3, was approximated by a piecewise linear function, as indicated. It is assumed that the fraction $F(D)$ is also stationary, and thus is independent of t.

Given the independence assumption, we have that

$$P\{\text{the peak on day } t \text{ falls within } D \text{ of the yearly peak}\} = F(D)w_t,$$

and (2.6)

$$W_T F(D) = E[\ \#\ \text{remaining peak days falling within } D \text{ of the yearly peak}].$$

Planning Problem Solution

Given the distributions above, a derivation is given in Appendix A that specifies k, the number of interruptions required, as a function of the total interruptible load and the number of peak days in the contract period. For planning, the formula for k requires the selection of the probability α that a peak which should have been curtailed is missed due to forecast error. The exact formula for the planning relationship k cannot be solved in closed form, but is shown graphically in Figure 4. This figure shows the function

$$k = k(y, W_T), \tag{2.7}$$

where $y = z_\alpha + L/\sigma_2$ and $\Phi(z_\alpha) = 1-\alpha$

determines z_α, for given confidence level α. For example, α=10% yields $z_\alpha = 1.28$ and for 100 MW of interruptible load and W_T=60, we have $y = 1.28 + 100/\sigma_2 = 1.76$ and thus $k = k(y, W_T) = 33$ interruptions from Figure 4. It can be seen from Figure 4 that increasing the interruptible load L will increase k the number of interruptions required. Similarly, decreasing α, the probability of missing one of the desired peaks, increases z_α which increases k.

Dispatching Problem Solution

The dispatching decision at time T for given k is to choose an optimal threshold H. This decision is independent of L and α. The dispatching threshold H is determined by

$$H = \mu - y\sigma_2. \tag{2.8}$$

where y is determined by looking up k on the vertical axis of Figure 4 and then using the appropriate W_T curve to obtain $y = y(k,W_T)$. Given a load L, it follows from the inverse of (2.7) that the probability of α of missing a peak L Megawatts below the annual peak satisfies

$$z_\alpha = y(k,W_T) - L/\sigma_2, \quad \text{with } \alpha = 1 - \Phi(z_\alpha). \tag{2.9}$$

Thus given k and T, the z_α and the corresponding probability α are adjusted depending on the interruptible load L.

Dispatching Procedure and Simulation

The dispatching methodology discussed above was implemented in a spreadsheet, illustrated in Figure 6, for use by the NEES system dispatcher. The user entries on the spreadsheet are shown in boxes and were coded as unprotected cells. The calculation portion of the spreadsheet is not shown. To begin the spreadsheet program operation, the user specifies the initial settings shown in the first square for Yearly Peak Forecast, Standard Deviation (of Forecast Error) and Total Interruptible Load. The user also enters the number of interruptions permitted (30 in this case) and then sequentially enters the Daily Forecast for each day. The spreadsheet calculates an updated Threshold each day, based on (2.8), with y determined as described in Appendix A. The W_T values are reduced each day by the individual daily weight w_t. The number of interruptions k is reduced as interruptions occur, based on the 0,1 values in the Int? column.

To test the methodology, the spreadsheet was applied retroactively to the 1988 data in a simulated day to day fashion. The results, as one might expect, depend crucially on how close the annual peak forecast μ is to the actual annual peak. For the data tested, the actual annual peak was 4280 MW. The simulation is illustrated in Figure 5, for initial annual peak estimates of 4100 MW, 4300 MW and 4500 MW, respectively. The 4100 initial estimate caused too many peaks in December and January to be interrupted, as indicated. When peaks higher than 4100 MW were observed, the forecasted annual peak was updated to the largest observed value, which caused the large jump in August. Smaller upward jumps occur in other cases, when the ratio k/W_T is reduced due to a rapid decrease in the remaining interruptions. This "self-correcting" characteristic of the threshold calculation prevents the remaining interruptions from being exhausted before the end of the year. For the 4100 MW forecast case, it also caused some of the large peaks occurring in later August to be missed because the number of remaining interruptions was too small.

3. Designing and Dispatching Multiple Block Contracts

In this section, we extend the results presented so far to interruptible loads that are segmented into multiple blocks, which are curtailed in a prescribed priority order as needed. The analytical development assumes that the amount of load curtailed is a continuous variable ℓ, ranging from zero to the total interruptible load L. We also show how this is implemented for the case of discrete blocks. The index ℓ is defined in such a way that $\ell=0$ is the first increment curtailed, and $\ell=L$ is the last. For a fixed value of α, a curtailment policy is equivalent to a set of thresholds $\{h_\alpha(\ell,L)\}$, which are monotonically increasing in ℓ, for any given value of α. The load increment ℓ is curtailed on day t if and only if the forecast $F_t \geq h_\alpha(\ell,L)$. With a single block of size L, as in Section 2, whenever the first increment $\ell=0$ is to be curtailed, the entire block will be curtailed. Thus $h_\alpha(0,L) = H_\alpha(L)$.

For analytical convenience, we derive planning and dispatching policies only for the case in which the same α value is used for all load increments ℓ. This leads to a simple form for the threshold policy, but is optimal only under certain circumstances. A sufficient condition is that the marginal net benefit of each KW of annual peak reduction is constant in the range $[A_0-L, A_0]$, where L is the total interruptible load. In other situations, it is possible that the optimal α could vary with ℓ, but we have not analyzed this case.

Given the forecasting error distributions used in this paper and fixed α for all load increments, we have

Lemma 3.1. The thresholds $h_\alpha(\ell,L) = H_\alpha(L-\ell) = H_\alpha(L) + \ell.$ (3.1)

Proof: Focusing on the load increment ℓ, α corresponds to the probability that an actual peak of $A_0-(L-\ell)$ is not curtailed, where A_0 is the actual annual peak. Given α, the curtailments of load increment ℓ should be the same as if we had a single block of size $L-\ell$ and curtailed according to the policy in Section 2. It is shown in Appendix A that curtailment occurs if and only if the sum of the forecast errors E_t+E_0 is such that $A_0-(L-\ell)+E_t+E_0 \leq H_\alpha(L-\ell)$. Thus it follows that

$$\alpha = P\{A_0-(L-\ell)+E_t+E_0 \leq H_\alpha(L-\ell)\} = P\{A_0-(L-\ell)+E_t+E_0 \leq h_\alpha(\ell,L)\},$$

while $H_\alpha(L-\ell) = H_\alpha(L) + \ell$ follows from (A.3). QED.

Lemma 3.1 simplifies the form of the optimal policy considerably, because it says in effect that the thresholds $\{h_\alpha(\ell,L)\}$ are implemented by curtailing down to the threshold $h_\alpha(0,L) = H_\alpha(L)$. That is, for any forecast such that $F_t > H_\alpha(L)$, the curtailed load should be the difference $\ell = F_t-H_\alpha(L)$.

The policy for each load increment ℓ is the same as if the total interruptible load were equal to $L-\ell$. Thus, the function $k=k_\beta(H_\alpha(L-\ell),T)$ defined in (A.12) gives the number of interruptions k such that β is the probability of running out of interruptions prematurely when the threshold $H_\alpha(L-\ell) = h_\alpha(\ell,L)$ is used for the remainder of the contract at time T. As discussed in the Appendix, the relationship $\beta=\alpha/2$, which approximates the optimal tradeoff between missing a peak and exhausting the available interruptions reduces (A.12) to (A.19), which holds for all values of ℓ. To minimize the probability that an actual peak $A_0-(L-\ell)$ is not curtailed, we use the relationship $k=k(y(\ell))$ in (A.19), which is plotted in Figure 4, where

$$y(\ell) = z_\alpha + (L-\ell)/\sigma_2. \tag{3.2}$$

This solves both the planning and dispatching problems is a manner analogous to the one block case with interruptible load $L-\ell$.

We have that $k(\ell) = k(y(\ell),W_T)$ is the number of times that load increment ℓ

should be curtailed with W_T peak days remaining. For dispatching any load increment ℓ with given k and W_T, we determine a corresponding y value from Figure 4 and then dispatch it according to the threshold value

$$H = \mu - y\sigma_2. \tag{3.3}$$

In the planning case, $k(\ell)$ can be chosen so that α is constant across all ℓ using (3.2). For the dispatching case (3.3), this is true at the beginning of the contract year, but is true in expectation only thereafter. That is, after some curtailments have occurred, the α value for the remainder of the contract year would differ from its initial values (as in the one block case), and would differ for different load increments.

Using Discrete Multiple Blocks

Discrete multiple blocks can be used in practice by selecting any staircase function that uniformly bounds the curve $k(\ell)$ from above as illustrated for two blocks in Figure 8. In this case, the interruptible load is segmented into two blocks of L_1 and L_2 units, where the subscripts denote order of interruption. Then k(0) interruptions are required of the low priority L_1 units, while $k(L_1) < k(0)$ interruptions are required for the higher priority L_2 units. For any given set of discrete blocks, an analogous procedure can be used to determine the number of interruptions required for each service priority block.

Figure 9 illustrates the appropriate procedure for dispatching multiple blocks. For the two blocks shown, L_2 is the load with higher priority (fewer curtailments per year), which has k_2 interruptions remaining, and L_1 is the lower priority load, which has k_1 interruptions remaining. For a given T value, we select the appropriate W_T curve and determine the two intersections with the horizontal lines drawn from k_1 and k_2. The y value for each block is determined by its right most edge. In general, the value y_i determines the threshold H_i for each block i from the relationship

$$H_i = \mu - y_i\sigma_2, \tag{3.4}$$

where the same procedure is used for any number of blocks.

At the start of the contract year, all the thresholds will satisfy

$$H_i = h_\alpha(\ell_i, L) \text{ where } \ell_i = \sum_{\{j \text{ interrupted before } i\}} L_j. \qquad (3.5)$$

As interruptions are used, some priority levels will become "out of balance." However, when a priority level has been interrupted too often, its threshold will be set disproportionately high based on (3.4), so that the imbalances will tend to correct themselves as the contract year progresses.

Determining a Price Schedule

In this section, we analyze the relative pricing of service priorities. We assume that the total interruption losses for each customer load type are linear in the number of interruptions per year, with the cost per interruption varying over the customer population. The distribution of interruption costs defines a demand function

$v(x)$ = the value such that at least x units of load have a cost per interruption which is $\leq v(x)$.
= credit per interruption which will elicit x units of interruptible load.

The electric utility must know the function $v(x)$ for the customer population, but need not know the interruption cost of any particular load unit in order to design the tariff. Economic efficiency dictates that a lower priority block should consist of load units of lesser interruption cost than a higher priority block. Pricing should induce customers to assign load units to service priorities in this manner, while the contracts for the blocks must also be consistent with the design considerations discussed in the previous section.

These goals can be achieved by a nonlinear credit schedule, in which the credit given to each customer is linear in the subscribed load and the credit per load unit is an increasing concave function of the allowable number of interruptions, illustrated in Figure 10. Since the function $k(\ell)$ specifies the number of required interruptions with ℓ or more curtailed units of load, its inverse $\ell(k)$ denotes the load ℓ that must be available for curtailment during the k most severe interruptions. This can be induced by setting the incremental credit $c(k)$ per load unit for the k^{th} interruption so that

$$c(k) = v(\ell(k)) \quad \text{for } k \geq k(L), \quad \text{and } c(k) = v(L) \text{ for } k \leq k(L). \qquad (3.6)$$

With (3.6), the number of interruptions k selected for each load unit $\ell(k)$ is such that the incremental credit equals the incremental interruption loss, and thus is incentive compatible. Further, since $v(\ell)$ is decreasing in ℓ, the marginal credits are positive and decreasing, thus establishing concavity. Finally, since $\ell(k)$ was used to determine the credit level for k interruptions, the number of subscriptions in each category will be consistent with (A.19) and (3.2).

Discrete Block Pricing

For discrete blocks, the above pricing approach would set the credit per interruption in each block to the value corresponding to the most costly load unit in that block. As shown in Figure 11, the first block (low priority) with L_1 units will be interrupted k_1 times while the second block (high priority) with L_2 units will be interrupted k_2 times. The incremental credit should be $v(L_1+L_2)$ for the first k_2 interruptions and $v(L_1)$ for the incremental k_1-k_2 interruptions. The total credit has a concave piecewise linear shape with the vertices corresponding to the offered contracts. This contract structure will induce the L_1 units with the lowest interruption losses to select the low priority contract and L_2 units with the next higher interruption losses to select the high priority contract. Generalizing to the case of n blocks with respective loads of L_1, ..., L_n and number of interruptions $k_1 \geq k_2 \geq ... \geq k_n$ will result in a menu of contracts with

$$C_n = k_n v\left(\sum_{j=1}^{n} L_j\right)$$

and

$$C_i = C_{i+1} + (k_i - k_{i+1})\, v\left(\sum_{j=1}^{i} L_j\right), \quad \text{for } i = 1,...,n-1.$$

where credit C_i is offered for k_i interruptions.

In the one block case, the utility must pay for each incremental interruption an amount that equals the highest loss incurred by any unit of load sustaining that interruption. Thus, the credits paid by the utility substantially exceed the cumulative interruption losses sustained by customers as illustrated in Figure 12. This "overpayment" in turn reduces the utility's incentive to enlist interruptible load, which reduces the social efficiency of the system. Multiple discrete blocks can significantly reduce the "overpayment" and thus provide incentives to increase social surplus. Figure 13 illustrates the saving for each of the k_1-k_2 interruptions in which the high priority L_2 units are not curtailed in the two block case. The area A_1 represents the saved social cost of the spared curtailments, while the area A_1+A_2 represents the savings to the utility based on the original credit for the one block dispatch. The utility achieves an additional savings A_3 due to the fact that the incremental k_1-k_2 interruptions can be compensated at a lower rate. Thus, the total social savings due to blocking is $(k_1-k_2)A_1$ while the credit savings to the utility is $(k_1-k_2)(A_1+A_2+A_3)$. These savings clearly increase with the number of blocks.

4. Conclusion

This analysis accomplished two main objectives. First, we developed an operational tool for dispatching a single block of interruptible load. This solved a practical problem faced by the New England Electric System and the spreadsheet solution described in this paper is currently in experimental use by the System Dispatcher. Our second objective was to develop insights and practical tools to aid the electric utility in the expansion of its interruptible service program. We developed a methodology for evaluating the costs and benefits of new service options for curtailable service and determining its most efficient features.

The planning methodology was useful in demonstrating to NEES that, given their current subscribers and demand patterns, there is limited benefit in enlisting more interruptions per year from existing CIS subscribers. Instead, the focus should be on enlisting new CIS subscribers, or additional load from current subscribers, by offering service options with fewer interruptions per year to induce additional subscription. Our analysis also showed how the number of required interruptions increases, as the amount of interruptible load increases in the single block case. As interruptible load is added, segmenting the load into blocks that are dispatched according to a priority order becomes more economically efficient and requires lower credit payments from the utility.

We have also shown in this paper how our dispatching algorithm can be extended to the multiple block case, but this generalization has not been implemented currently. For the multiple block case, we have also presented a methodology for pricing a menu of service options with different numbers of interruptions during the contract period. The pricing methodology requires knowledge only of the distribution of interruption losses in the population and is based on some simplifying assumptions. The resulting menu of service options induces self-selection by customers. That is, service priorities are selected for load units in a way that produces a socially efficient resource allocation and the desired contract mix for the utility.

5. References

Chao, Hung-po, Robert B. Wilson (1976), "Priority Service: Pricing, Investment and Market Organization," American Economic Review, Vol. 77, No. 5, , pp. 899-916.

Chao, Hung-Po, Shmuel S. Oren, Stephen A. Smith and Robert B. Wilson (1986), "Multilevel Demand Subscription Pricing for Electric Power," Energy Economics, pp. 199-217.

----------------- (1987), "Selected Papers on Priority Service Methods," EPRI Report P-5350. Electric Power Research Institute (August)

-----------------(1989), "Priority Service: Market Structure and Competition," The Energy Journal, Vol 9, pp. 77-103.

Crew, M.A., P.R. Kleindorfer, "Reliability and Public Utility Pricing," The American Economic Review, Vol. 68, No. 1, 1978, pp. 31-40.

Feller, W. (1957) An Introduction to Probability Theory and its Applications, Vol. 1, John Wiley and Sons, New York.

Fisinger, J. (1980), "Peak Load Pricing and Rationing Policies," Journal of Economics, Vol 40. No. 2, pp. 169-182.

Marchand, Maurice G. (1974), "Pricing Power Supplied on an Interruptible Basis," European Economic Review, Vol 5, pp.263-274

Oren, Shmuel, Joseph A. Doucet (1990), "Interruption Insurance for Generation and Distribution of Electric Power," Journal of Regulatory Economics, (to appear in March 1990)

Smith, Stephen A. (1989), "Efficient Menu Structure for Pricing Interruptible Electric Power Service," Journal of Regulatory Economics, Vol 1, No. 3, pp. 203-224.

Telson, M. (1975), "The Economics of Alternative Levels of Reliability for Electric Power Generation Systems," Bell Journal of Economics, Vol. 6 (Autumn), pp. 679-694.

Tschirhart, J. and F. Jen (1979), "Behavior of a Monopoly Offering Interruptible Service," Bell Journal of Economics, Vol. 10 (spring), pp 244-258.

Viswanathan, N. and E. Tse (1989), "Monopolistic Provision of Congested Service with Incentive Based Allocation of Priorities," International Economic Review, No. 1 pp.

153-174.

Vickery, W. (1971), "Responsive Pricing of Public Utility Services," Bell Journal of Economics and Management Science, Vol 2. pp. 337-346.

Wharton J. B.(1988), "Customer Response and Appropriate Credits for Interruptible Rates: NEES Cooperative Interruptible Service Program," Working Paper, New England Power Service Company.

Wilson, Robert B.(1989), "Efficient and Competitive Rationing," Econometrica, Vol. 57, No. 1, pp. 1-40.

Wilson, Robert B. (1989), "Ramsey Pricing of Priority Service," Journal of Regulatory Economics, Vol 1, No. 3, pp.189-202.

6. Appendix A

The models derived in this appendix are used to develop the relationships illustrated in Figure 4. Suppose that our goal is to interrupt a peak if it falls within L of the annual peak A_0. Let us consider a threshold policy in which an interruption is called for if the daily forecast exceeds the threshold H. For any given H, there is some probability that a peak which is actually A_0-L will be missed due to forecast error, namely

$$\alpha = P\{\text{the forecast} < H \text{ on a day with peak } A_0-L\}$$

Since the forecast is increasing in the actual peak size, this is the worst case, for peak in the range $[A_0-L, A_0]$. In general, α could also depend on the time of the year T, but we have assumed that annual peak forecast errors are stationary throughout the year.

Conversely, for given α and L, we can determine a threshold $H_\alpha(L)$ such that a peak of size A_0-L is missed with probability α based on (2.1) and (2.2). In general, the forecast on a day with peak A_0-L satisfies

$$F_t = A_0 - L + E_t + E_0 \tag{A.1}$$

where F_t = daily forecast on day t

E_t = error in the daily forecast for day t

E_0 = error in forecasting the annual peak.

Now define the random variable $Z_t = A_0 + E_t + E_0$. If we assume that E_0 and E_t are independent, it follows from the additive property of Normal distributions that

$$P\{Z_t \leq z\} = \Phi\left(\frac{z-\mu}{\sigma_2}\right), \tag{A.2}$$

where $\sigma_2 = \sqrt{\sigma^2+\sigma_1{}^2}$, based on the distributions defined in Section 2.

The probability α that the peak A_0-L is missed equals the probability that the forecast on that day is less than H. Thus, for the threshold $H_\alpha(L)$, we must have $\alpha = P\{Z_t-L \leq H_\alpha(L)\}$ from (A.1). Since Z_t is normally distributed, we can obtain a corresponding z_α value from a Normal table and obtain $H_\alpha(L)$ from

$$H_\alpha(L) = \mu - L - z_\alpha\sigma_2. \tag{A.3}$$

In addition to missing a peak due to forecast error, it could also be missed if the remaining interruptions are exhausted before it occurred. Thus, it is important to consider the probability

$$\beta = \beta(H,k,W_T) = P\{\, k \text{ interruptions are used up, given threshold } H \text{ and } W_T \text{ peak days remaining}\}.$$

For any day t and given H, an interruption occurs if and only if $F_t \geq H$. To consider the probability that this occurs, we need to consider all peaks that may trigger an interruption. Let D_t be the random variable corresponding to the difference between the peak on day t and the annual peak A_0. Using the random variables E_t and E_0 discussed above, the threshold H is exceeded by the forecast on day t if and only if

$$A_0 - D_t + E_t + E_0 \geq H \quad \text{or} \quad D_t \leq Z_t - H, \tag{A.4}$$

using the random variable Z_t defined above.

For any value D, let us define an indicator random variable

$$x_t(D) = \begin{cases} 1 & \text{if day } t \text{ falls within } D \text{ of the yearly peak} \\ 0 & \text{otherwise,} \end{cases}$$

with

$$P\{x_t(D)=1\} = F(D)w_t \tag{A.5}$$

based on (2.5). The probability that the forecast on day t is greater than H is

$$P\{x_t(Z_t-H)=1\} = w_t\int_H^\infty F(z-H)\phi\left(\frac{z-\mu}{\sigma_2}\right)\frac{dz}{\sigma_2} = w_t\int_0^\infty F(\sigma_2 z)\phi\left(z-y\right)dz = w_t I(y), \tag{A.6}$$

where $y = (\mu-H)/\sigma_2$.

The function $I(y)$ is the expected fraction of days whose forecasts are within y standard deviations of μ, the forecasted annual peak, or equivalently, the expected fraction of peak days whose forecasts are above the threshold H. The variable $x_t(Z_t-H)$ is a Bernoulli random variable with

$$E[x_t(Z_t-H)] = w_t I(y) \text{ and } Var[x_t(Z_t-H)] = w_t I(y) - w_t^2 I^2(y). \tag{A.7}$$

Given the current day is T, the number of remaining days whose forecast exceeds H is the random variable

$$S_T(H) = \sum_{t=T}^{365} x_t(Z_t-H). \tag{A.8}$$

It can be shown that the sum $S_T(H)$ satisfies the Central Limit Theorem. This follows from the fact that it is a sum of independent random variables which are uniformly bounded in absolute value [see Feller(1957), pp.238-9]. Thus we have the approximate relationship

$$P\{S_T(H) \le k\} \approx \Phi\left(\frac{k-W_T I(y)}{s(y,T)}\right), \tag{A.9}$$

where $y = (\mu-H)/\sigma_2$

$W_T I(y) = E[\text{number of remaining forecasts greater than } H]$

$s(y,T) = \sqrt{W_T I(y) - I^2(y)\sigma_T^2}$ = standard deviation of the number of remaining forecasts greater than H

and $W_T = \sum_{t=T}^{365} w_t$ and $\sigma_T^2 = \sum_{t=T}^{365} w_t^2$.

Since β(H,k,T) is the probability that the remaining interruptions are exhausted, it is clear, given there are k interruptions remaining on day T, that

$$1-\beta(H,k,T) = P\{S_T(H) \leq k\}. \tag{A.10}$$

Conversely, for given β, H and T, we can determine the number of interruptions $k_\beta(H,T)$ that are required. Using β, we obtain a value z_β from a standard Normal table, which based on (A.10) must satisfy

$$z_\beta = \frac{k - W_T I(y)}{s(y,T)}. \tag{A.11}$$

We then solve (A.11) to obtain a formula for $k_\beta(H,T)$

$$k = k_\beta(H,T) = W_T I(y) + z_\beta s(y,T). \tag{A.12}$$

In (A.12), y is the number of standard deviations that H is away from the forecasted annual peak μ, i.e., $y = (\mu - H)/\sigma_2$. Given that the interruptible load is L, we can use (A.3) to substitute for H, which yields

$$y = z_\alpha + L/\sigma_2. \tag{A.13}$$

Solving the Dispatching Problem

For the dispatching problem with given k, the goal of the dispatching methodology reduces to one of attempting to interrupt the k highest remaining daily forecasts in the year. This can be seen intuitively because the optimal use of the k interruptions is to interrupt the k highest peaks. Since the errors in daily forecasts are time independent, this implies that the highest k forecasts are the most likely candidates for the k highest peaks. It is clear that the dispatching solution should depend on k and T, but not on L and, in effect, we solve the case $L = 0$.

For known k, (A.12) and (A.13) provide two equations for the three unknowns

y, z_α and z_β. Thus, one additional relationship is required to obtain an optimal value of y, which leads to an optimal threshold H for the dispatching problem. The solution is completed by determining a relationship between z_α and z_β that balances the probability of missing a peak due to forecasting error and missing a peak due to running out of remaining interruptions.

Let β equal the probability of running out of interruptions and α equal the probability that the annual peak A_0 is missed due to forecast error. Then the probability that the annual peak is missed can be expressed as

$$\begin{aligned} P\{\text{missing annual peak}\} = \; & \beta P\{A_0 \text{ occurs after running out} \mid \text{run out}\} + \\ & (1-\beta)P\{\text{forecast on day of } A_0 < H \mid \text{given that don't run out}\}. \end{aligned}$$

As a first order approximation, we assume independence permits the conditioning to be ignored in the two conditional probabilities above. We also take the expected value case in which the interruptions are used at exactly their average rate during the remainder of the year. This implies that β is the probability that the annual peak occurs after the interruptions are exhausted, given that they are exhausted. Thus, substituting these assumptions we have

$$P\{\text{miss annual peak}\} = (1-\beta)\alpha + \beta\beta = \alpha - \alpha\beta + \beta^2. \tag{A.14}$$

We then wish to minimize this expression with respect to the ratio $\alpha/\beta = r$. The probability can be rewritten as

$$\alpha - r\alpha^2 + r^2\alpha^2. \tag{A.15}$$

It can be verified from the first and second order necessary conditions that this expression has a unique minimum at $r = 1/2$, i.e. $\beta/\alpha = 1/2$, for any fixed value of α. Thus we have established a second relationship based on (A.15)

$$\beta = \alpha/2. \tag{A.16}$$

Using (A.16) with a standard Normal distribution table determines a trajectory of possible $(z_\alpha,\ z_\beta)$ pairs. Applying linear regression to the points shown gives the approximate relationship

$$z_\beta = 0.8022\ z_\alpha + 0.6272, \tag{A.17}$$

which provides an excellent fit, with an R^2 value of 0.997.

Substituting (A.17) into (A.11), we obtain

$$0.8022\ z_\alpha + 0.6272 \ = \ \frac{k - W_T I(y)}{s(y,T)}. \tag{A.18}$$

For the dispatching problem, we consider the case $L = 0$ so that (A.13) reduces to $y = z_\alpha$, which when substituted into (A.18) reduces to

$$k = W_T I(y) + s(y,T)[0.8022y + 0.6272]. \tag{A.19}$$

This implicit relationship can be used to obtain a tabular solution for y and thus the threshold H by substituting in various values of y and T and calculating the corresponding k. This function is shown graphically in Figure 4. To solve the dispatching problem for a given k, T, we determine $y = z_\alpha$ from Figure 4. Then the corresponding dispatching threshold H is

$$H = \mu - y\sigma_2, \tag{A.20}$$

which is equivalent to $H_\alpha(0)$ in (A.3).

Solving the Planning Problem

For the planning problem, we wish to determine the appropriate number of interruptions to allocate for the year. For a given interruptible load L, the planning problem objective is to determine how many interruptions are required so that the probability of missing a peak within L of the annual peak is α, given that they are dispatched in the manner discussed above.

The solution of the planning problem begins with the same relationship (A.19), which describes the solution of the dispatching problem. However, in this case, k is not specified, but any choice of y determines a dispatching threshold H from (A.20). To compute the probability that the dispatching threshold misses a peak that falls within L of the annual peak, we determine what z value is implied by the threshold H, by setting it equal to the righthand side of (A.3) to obtain

$$\mu - L - z_\alpha\sigma_2 = H = \mu - y\sigma_2. \tag{A.21}$$

That is, the probability that a peak within L of the annual peak is missed by a threshold $H = \mu - y\sigma_2$ is $\alpha = 1 - \Phi(z_\alpha)$, where z_α is determined from a Normal table. Thus to solve the planning problem for a given L and probability α, we set $y = z_\alpha + L/\sigma_2$ and obtain the appropriate k from (A.17).

In the spreadsheet implementation of the dispatching threshold, it was necessary to obtain approximate linear relationships for I(y) and s(y,T) to simplify (A.17). Using the numerical integration and applying regression to the results gave

$$I(y) \approx 0.0089 + 0.2306y, \text{ where } y = (\mu - H - L)/\sigma_2 \tag{A.22}$$

$$s(y,T) \approx \sqrt{W_T}\{0.2875 + 0.1464y\}. \tag{A.23}$$

The regression for s(y,T) used the T value corresponding to $W_T = 30$. The R^2 values in these regressions were both larger than 0.99. The fits of the functional forms in (A.22) and (A.23) are shown in Figure 7 for $W_T = 15$, 30 and 60.

Substituting in (A.22) and (A.23), (A.19) becomes a quadratic function to be solved for H. Using the quadratic formula, and selecting the positive root for y, we have

$$y = \frac{-B + \sqrt{B^2 - 4AC}}{2A}. \tag{A.24}$$

where $A = 0.1174$, $B = 0.2306\sqrt{W_T} + 0.3224$, $C = 0.1803 + \sqrt{W_T}[0.00887 - k/W_T]$. This formula was used for the spreadsheet in Figure 6.

Figure 1. Daily Peak Loads 1988

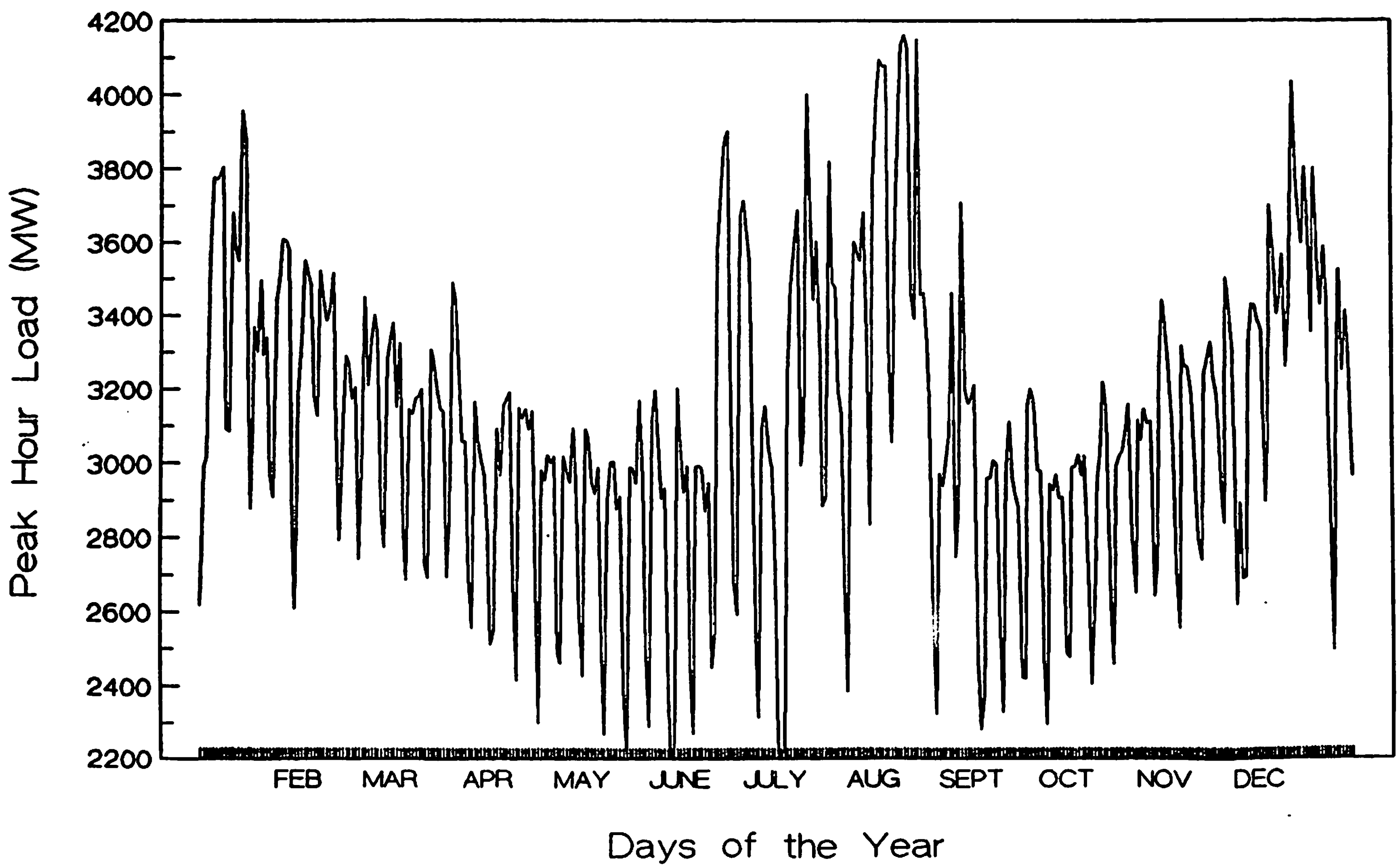


Figure 2. Distribution of Forecast Errors in Top Peaks

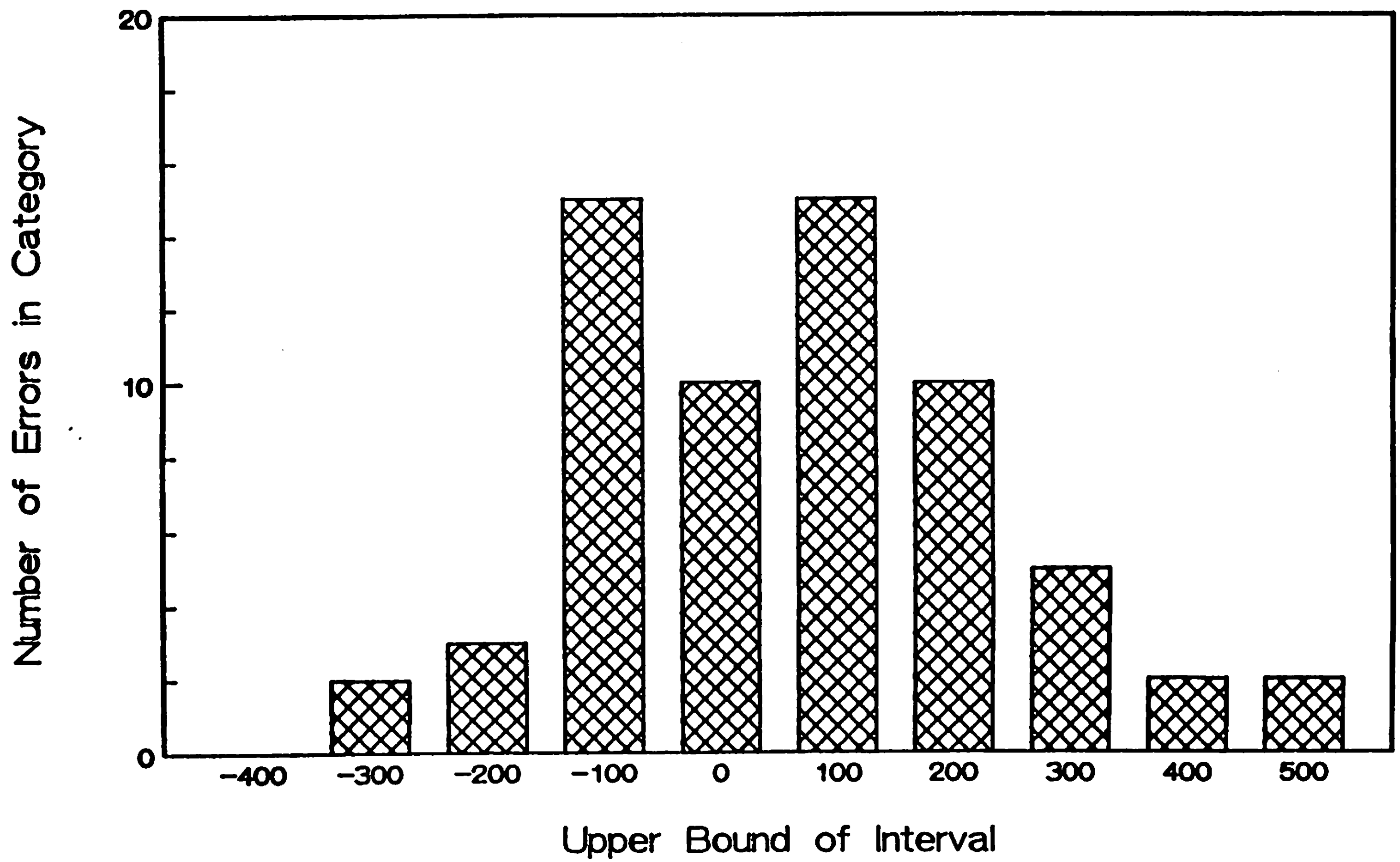
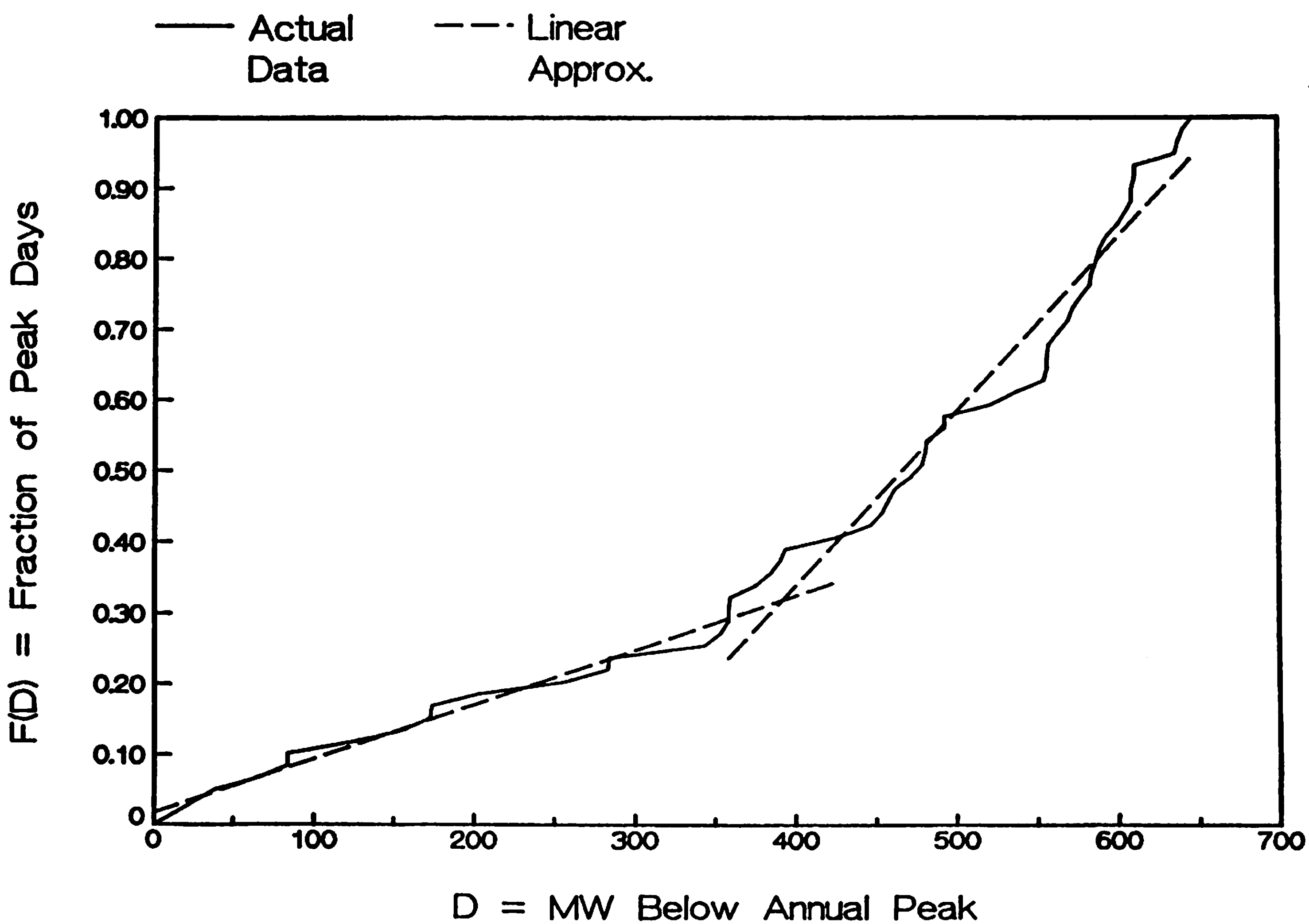


Figure 3. Fraction of Peak Days within D of the Annual Peak



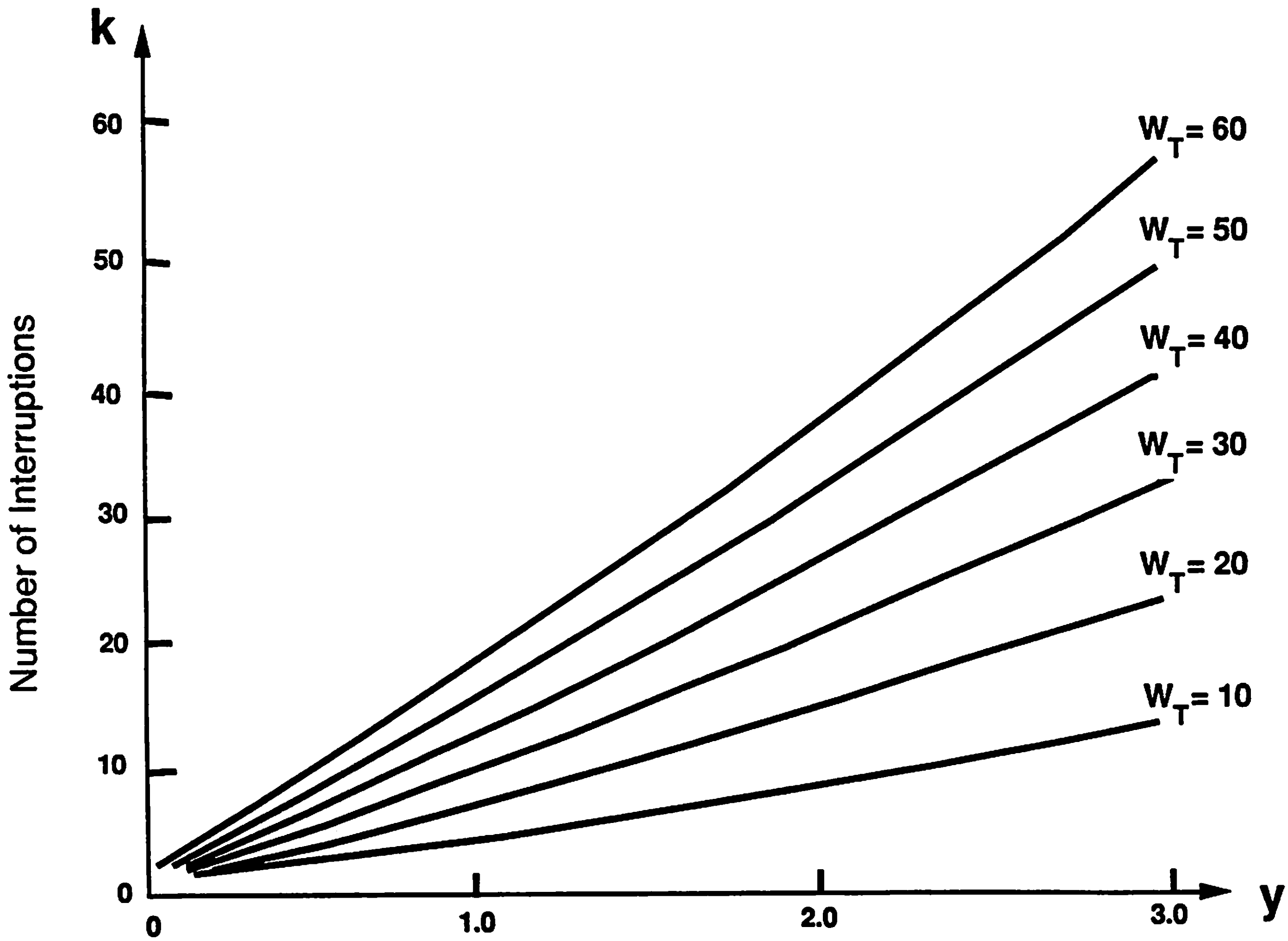


Figure 4: Number of Required Interruptions

Figure 5. Daily Forecasts and Threshold Patterns for Three Annual Peak Forecasts

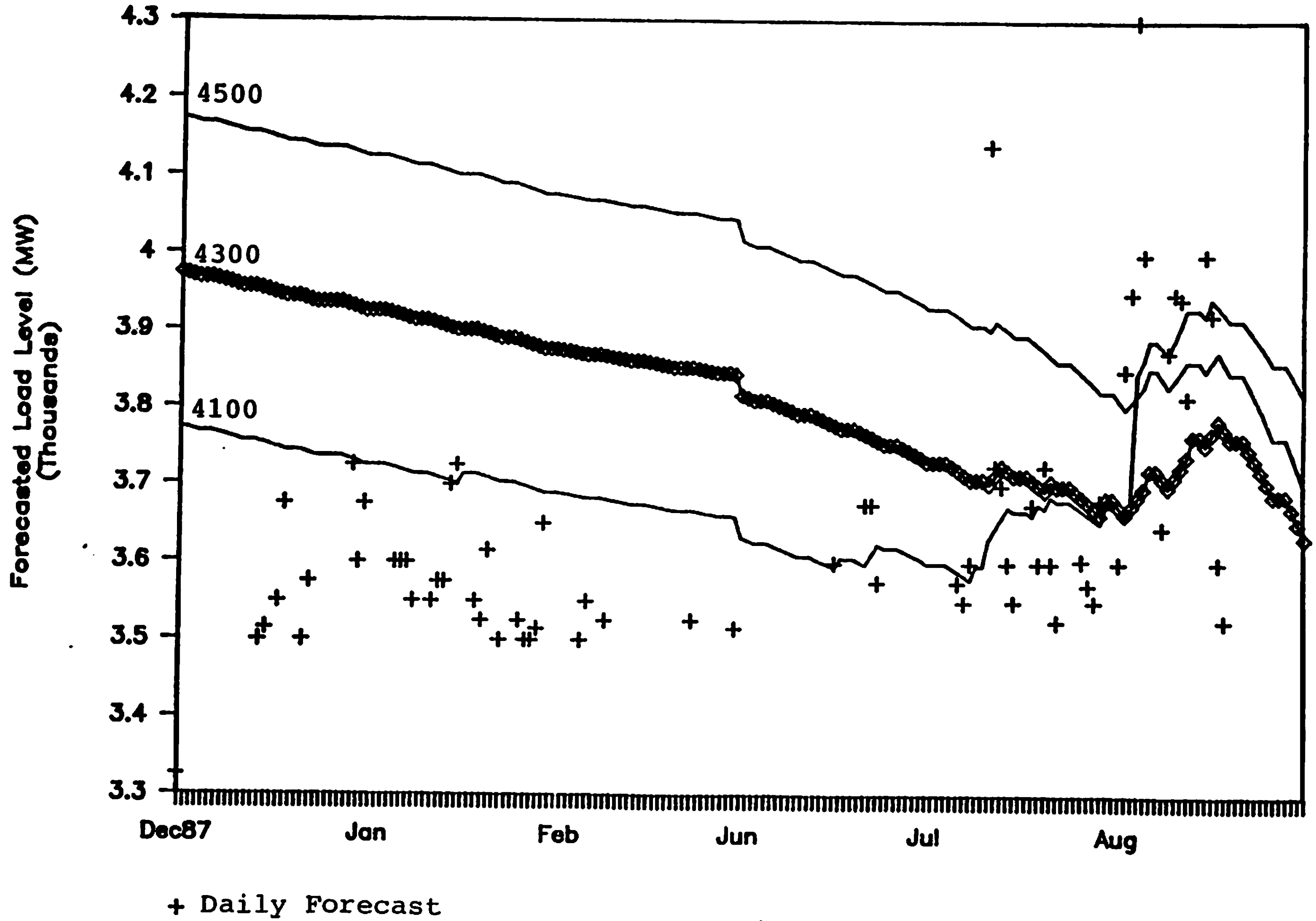


Figure 6. Spreadsheet Implementation

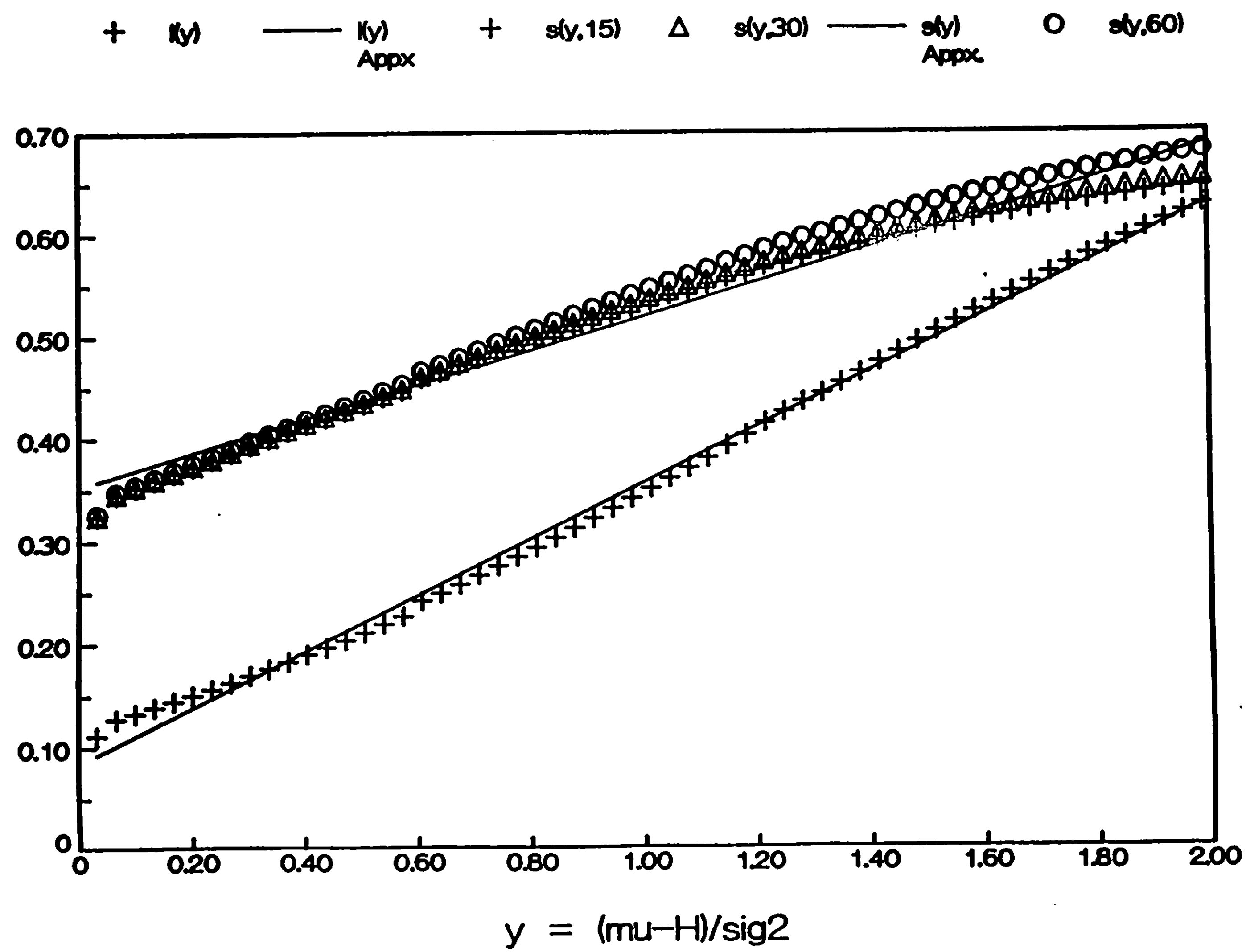
1988 Application of Threshold Model

Yearly Peak Forecast	4300	MW
Standard Deviation	205	MW
Interruptible Load	100	MW

16 Interr. Used to Date

	Month	Day	Day of Week	wt	60 WT	30 k	Daily Frcst	Thresh	Int?	Yrly Peak Estim
Dec87	12	1	2	0.455	60	30	3325	3874	0	4300
	12	2	3	0.455	59	30	3325	3872	0	4300
	12	3	4	0.455	59	30	3425	3870	0	4300
	12	4	5	0.455	58	30	3950	3867	1	4300
	12	5	6	0.000	58	29	2500	3878	0	4300
	12	6	7	0.000	58	29	2500	3878	0	4300
	12	7	1	0.455	58	29		3876		4300
	12	8	2	0.455	57	29		3874		4300
	12	9	3	0.455	57	29		3872		4300
	12	10	4	0.455	56	29		3869		4300
	12	11	5	0.455	56	29		3867		4300
	12	12	6	0.000	56	29		3867		4300
	12	13	7	0.000	56	29		3867		4300

Figure 7. Linear Approximations for I(y) and s(y,T)



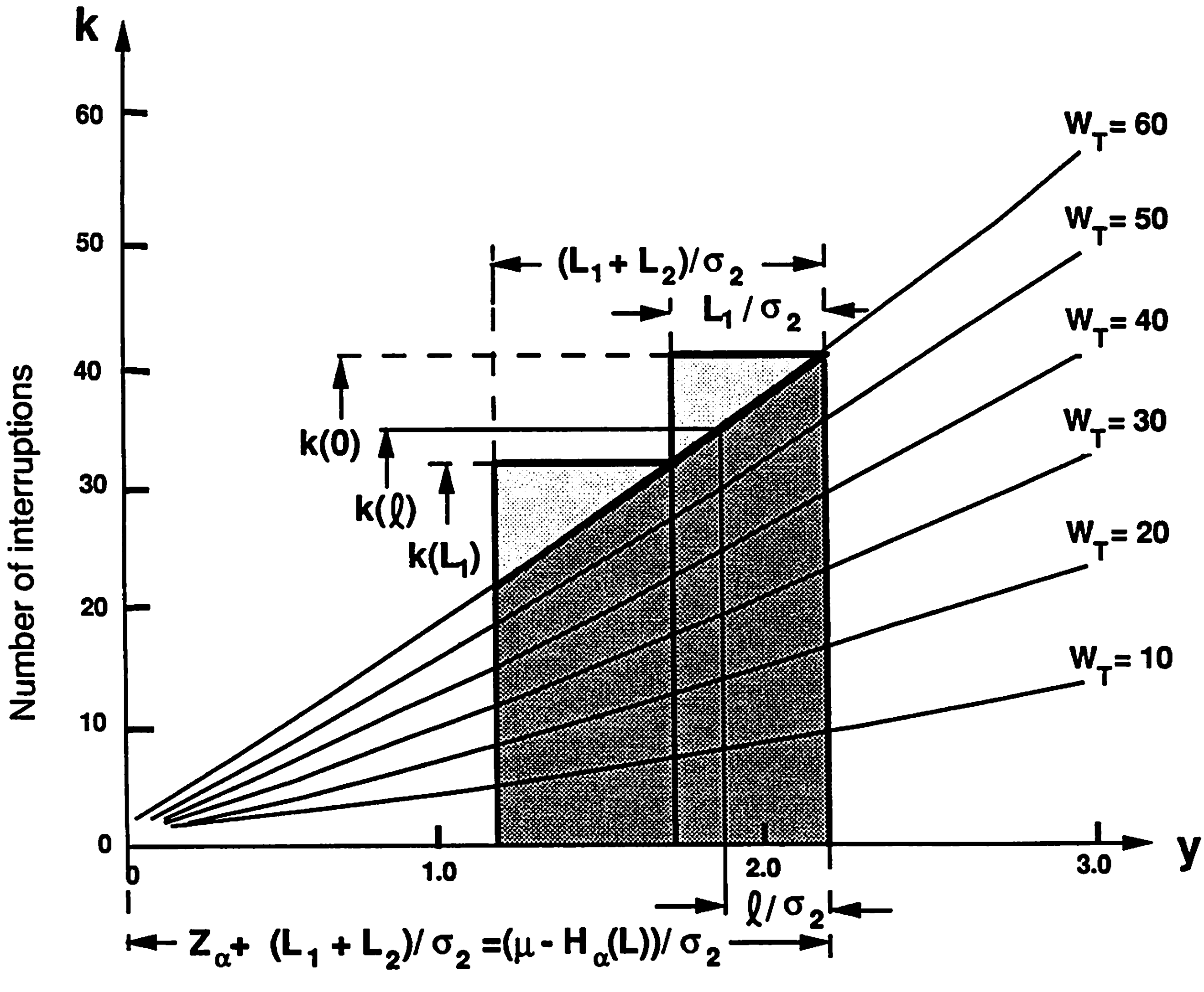


Figure 8: Number of Interruptions for Multiple Blocks

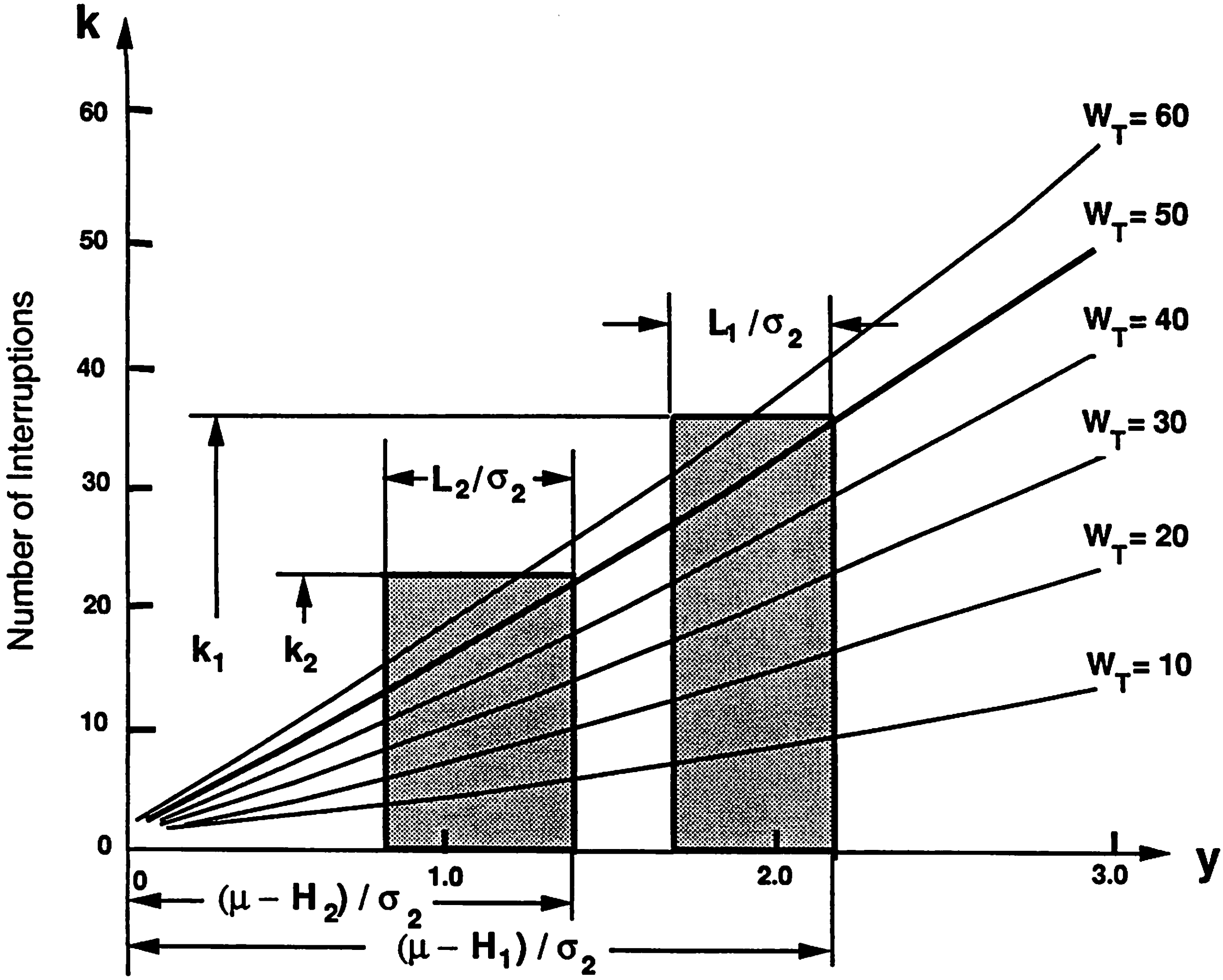


Figure 9: Determining the Thresholds in Multiblock Dispatch

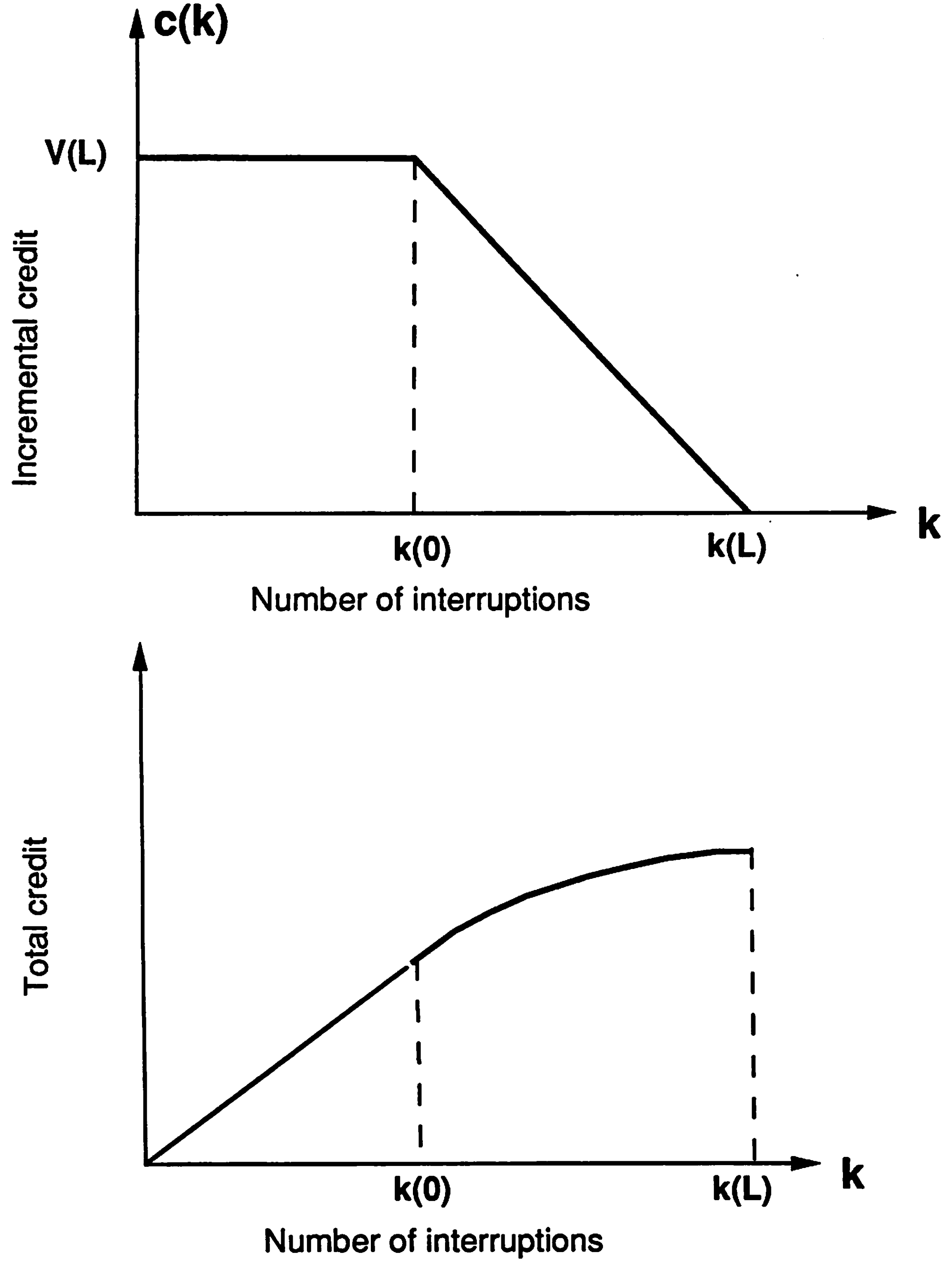


Figure 10: Credit Structure for Continuous Dispatch

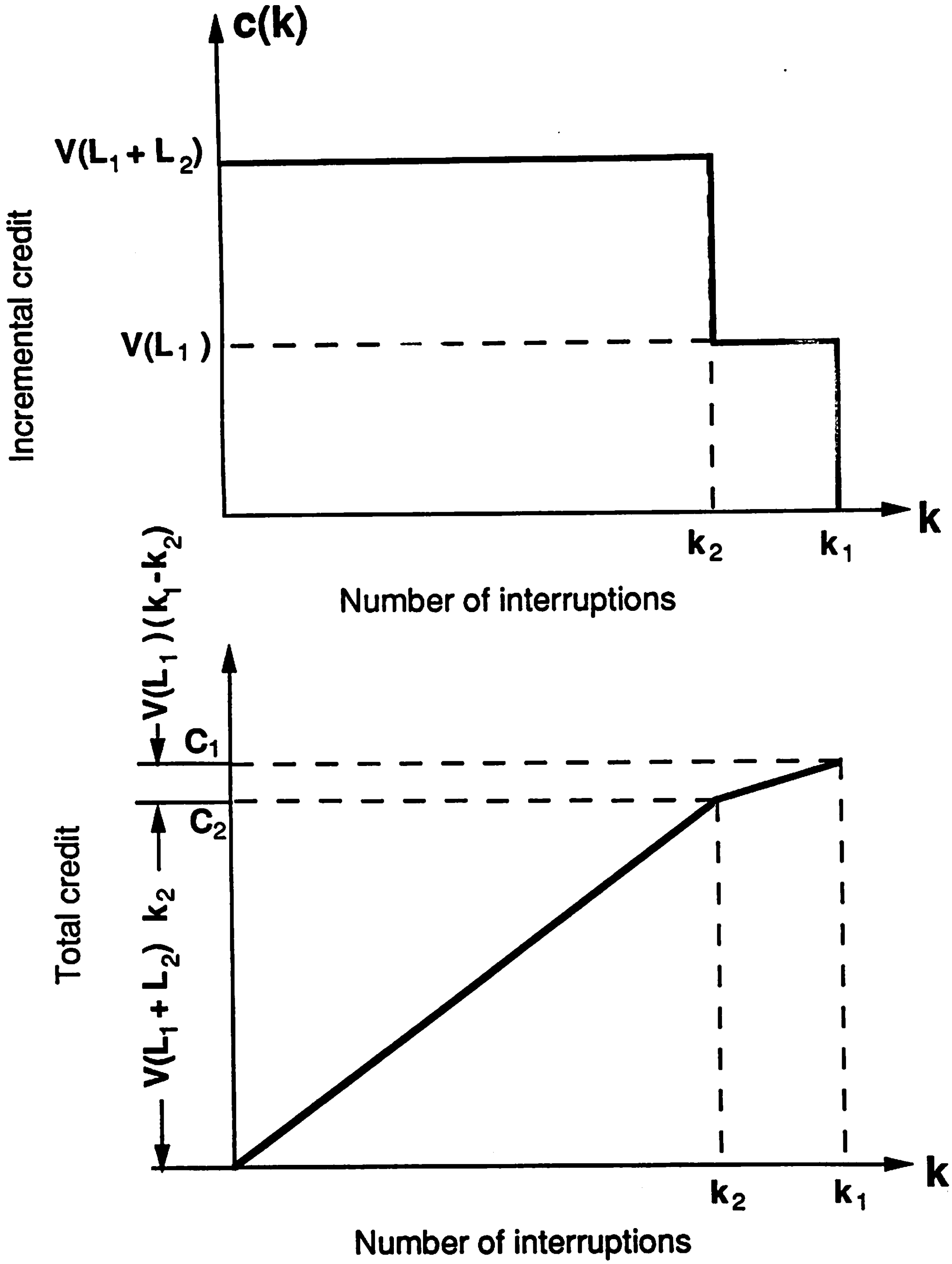


Figure 11: Credit Structure for Two Blocks

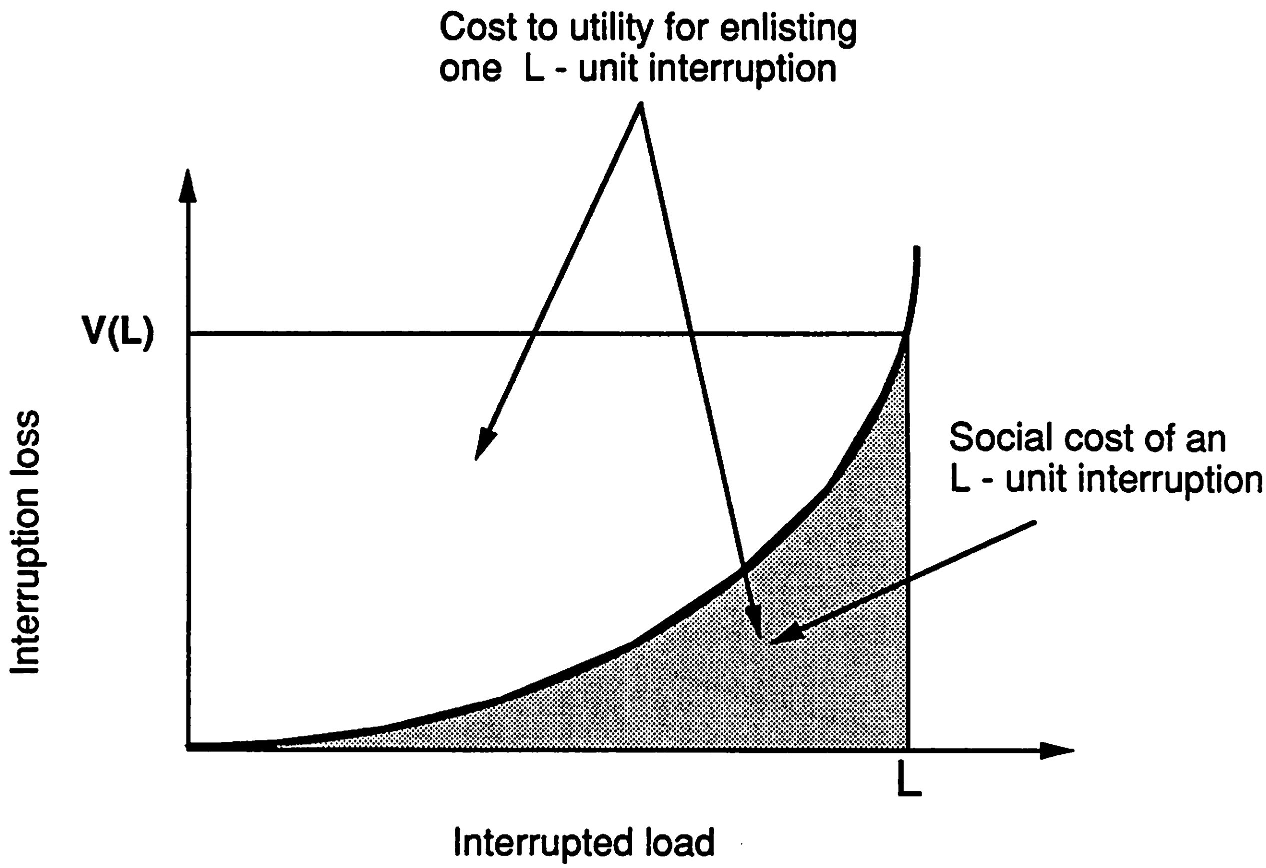


Figure 12: Cost to Utility vs. Social Cost per Interruption

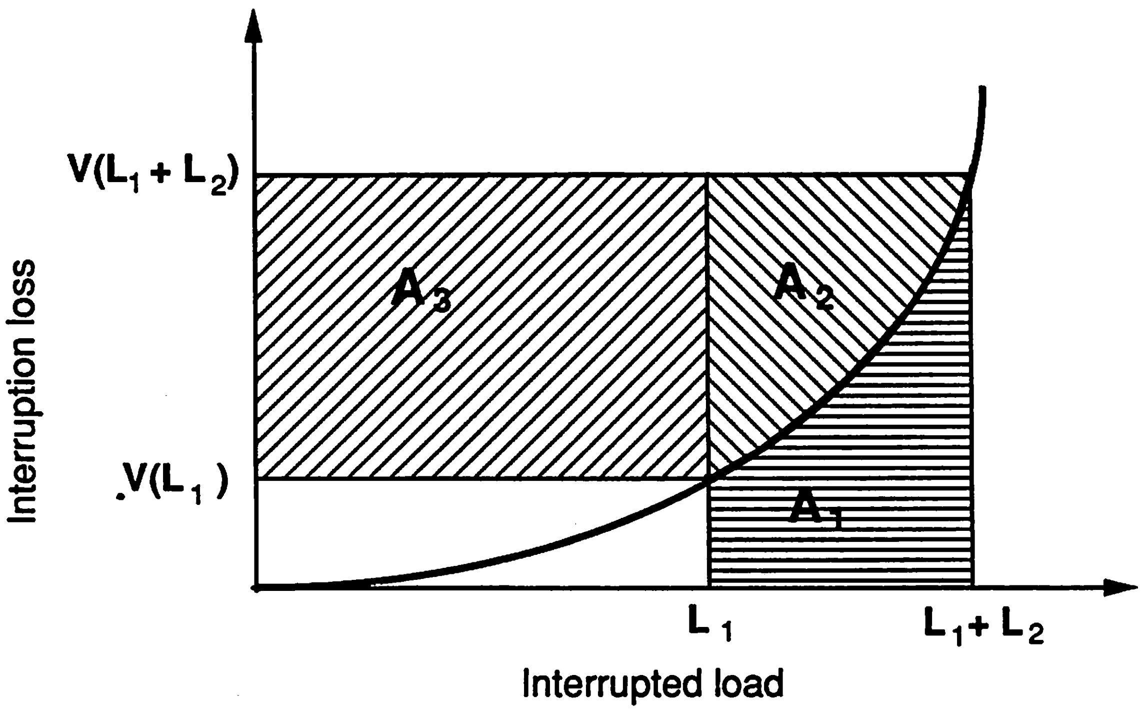


Figure 13: Savings Due to Blocking of the Interruptible Load