WORKING PAPER
ALFRED P. SLOAN SCHOOL OF MANAGEMENT

A Model for the Configuration of Incoming WATS Lines

by
Roger H. Blake
Stephen C. Graves
P. Clark Santos

WP #3134-90-MSA March 1990

MASSACHUSETTS
INSTITUTE OF TECHNOLOGY
50 MEMORIAL DRIVE
CAMBRIDGE, MASSACHUSETTS 02139

A Model for the Configuration of Incoming WATS Lines

by
Roger H. Blake
Stephen C. Graves
P. Clark Santos

WP #3134-90-MSA March 1990

A Model for the Configuration of Incoming WATS Lines

Roger H. Blake*
WearGuard
Director of Decision Support Systems
Longwater Drive
Norwell, MA 02061

Stephen C. Graves
Massachusetts Institute of Technology
Sloan School of Management
Room E53-390
Cambridge, MA 02139

P. Clark Santos
currently employed by
AT&T
Consumer Products
5 Woodhollow Road
Room 2J41
Parsippany, NJ 07054

January 1990

Abstract

WearGuard is a direct marketer and retailer of uniforms and work cloths, which relies primarily on phone orders for sales. For this purpose it maintains a series of toll-free "800-number" lines, known as WATS lines, to receive its incoming calls. These lines are of several types, where each type serves a different portion of the country and has a different usage fee. In this paper we determine how many of each type of WATS lines should be employed. After defining the problem more completely, we develop a queueing model to describe the system and a dynamic program to solve the configuration problem to optimality. The model has been applied to the problem by WearGuard since 1984. We present an example and examine the sensitivity of the solution to variations in various parameters. We validate the model by comparing the results of this model to other approximate models.

Key words: queueing application, overflow, queues, WATS lines.

*As of 1990, Manager of Strategic Services, Andersen Consulting, Boston, MA

1. BACKGROUND AND INTRODUCTION

In this paper we describe an application of queueing theory to the problem of determining the number of WATS lines for a phone-order operation. We developed the model in 1984 as part of the thesis project for one of us (Santos [5]), where the application to WearGuard served as the primary motivation. Since then, WearGuard has periodically used the model to evaluate and (re) configure the number and type of WATS lines in its telemarketing operations. Typically, the model is used once or twice a year, triggered by increases in call rates and by changes in the cost structure for calls. Recently, due to the introduction of a new type of service (unbanded T-1 lines) by its long-distance carrier, the crux of the problem has been eliminated for WearGuard. Nevertheless, the general structure of the problem remains for many telemarketing operations.

(a) WATS lines. AT&T Communications is the division within AT&T which is responsible for long distance services. In addition to the typical long distance service that most private customers use, AT&T Communications (as well as the other major long distance carriers) offers Wide Area Telephone Services (WATS) which are used by many industries. One type of WATS service allows a firm to receive calls from anywhere within a region at no cost to the caller, so-called toll-free "800" numbers. There are seven types of toll-free numbers, which depend upon the portion of the country the WATS line is to serve. For WearGuard, whose telemarketing base is in Massachusetts, the country is broken up into six zones plus the state of Massachusetts. For instance, Zone I consists of Eastern Pennsylvania, Eastern New York, New Jersey, and all of New England except Massachusetts; Zone II is Delaware, the District of Columbia, Ohio, Maryland, Western Pennsylvania, Western New York, Virginia, and West Virginia. Figure 1 shows the five zones. A Band I WATS line can be used for calls originating from Zone I, a Band II line for calls originating from Zones I and II, etc. There is also a Band VI line which can handle calls from Zone I through Zone V plus Alaska and Hawaii, and an additional service (Band IX) for calls coming only from the local state (Massachusetts). For the purpose of this study, however, these two bands will not be considered because the service of in-state calls represents a separate problem and the number of calls that come from Alaska and Hawaii do not justify the expense of maintaining the service to these states for WearGuard.

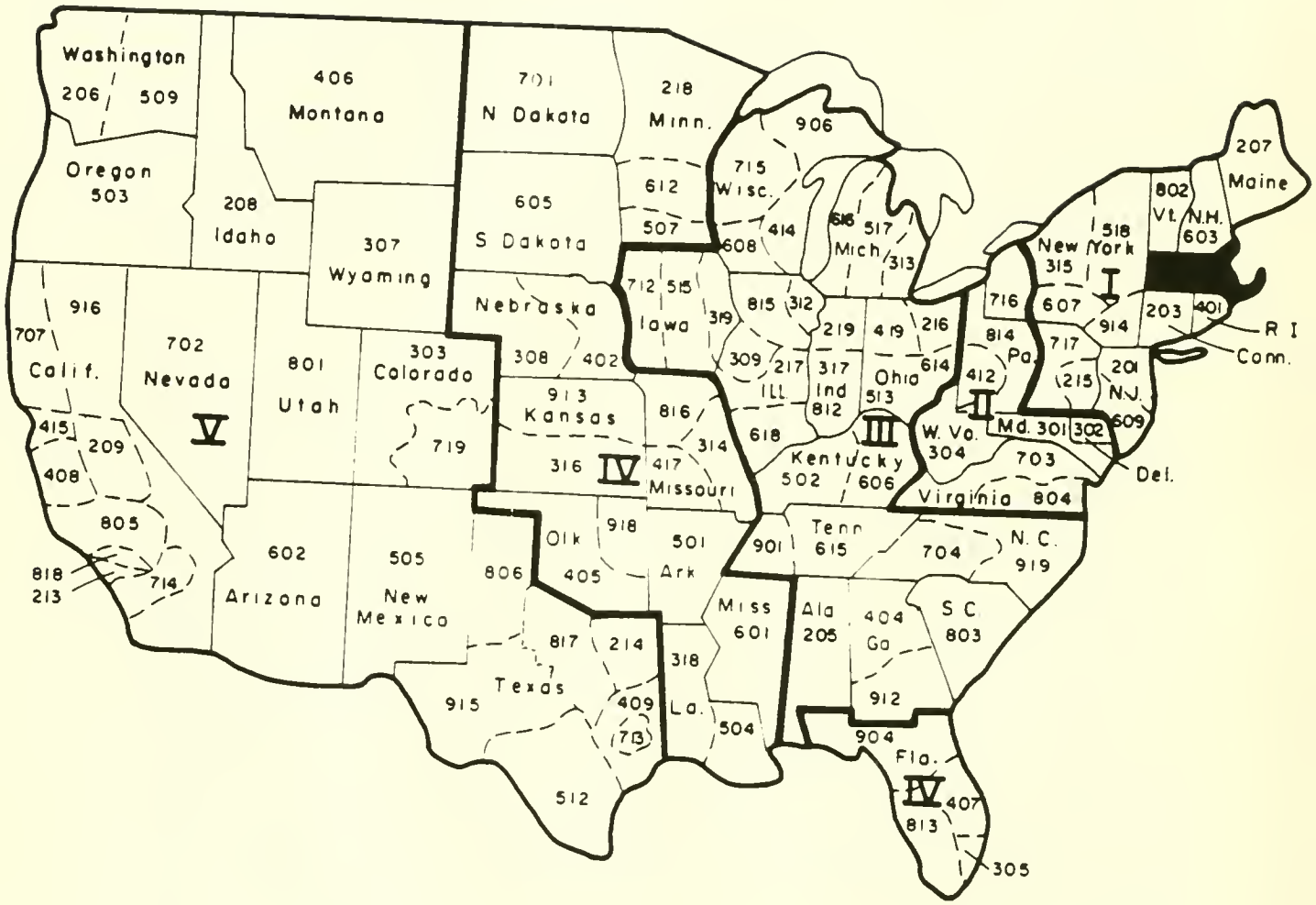


FIGURE 1: 800 SERVICE AND OUTSIDE WATS
SERVICE AREAS: MASSACHUSETTS

Nevertheless, it would be very easy to extend the model to include Band VI, whereas the intrastate lines, Band IX, must be treated separately.

In 1984 the cost of a WATS line consisted of a monthly access fee of $36.80 and an usage fee which depended upon the band, the total monthly usage of the lines, and the time of day during which calls are made (see Table 1). For example, if a user acquired a Band I line and used it for fourteen hours over the course of a month, he would be charged an access fee of $36.80 and an usage fee, which depends on when the fourteen hours were incurred. The usage charge would be $17.93 for each hour the phone was used during the business day, $12.51 for each hour the phone was used during the evening, and $8.54 for each hour of night and weekend usage. If, on the other hand, he used the phone for a total of sixteen hours, he would be charged $36.80 plus $16.36 for each hour of usage during the day, $11.79 for each hour in the evening, and $8.54 for each hour at night. When there are multiple lines of the same type, the usage rate depends on the average monthly usage over the set of lines.

Table 1: Hourly WATS Usage Rates (1984)

Band	Rate Period	0-15	15-40	40-80	Over 80
I	Day	17.93	16.36	14.82	13.13
	Evening	12.51	11.79	10.67	9.45
	Night/weekend	8.54	8.54	8.54	8.54
II	Day	18.88	17.26	15.63	13.83
	Evening	13.60	12.42	11.25	9.96
	Night/weekend	8.99	8.99	8.99	8.99
III	Day	19.58	17.88	16.20	14.33
	Evening	14.10	12.89	11.66	10.32
	Night/weekend	9.32	9.32	9.32	9.32
IV	Day	19.92	18.21	16.48	14.59
	Evening	14.34	13.12	11.87	10.51
	Night/weekend	9.49	9.49	9.49	9.49
V	Day	20.59	18.81	17.02	15.08
	Evening	14.82	13.54	12.26	10.86
	Night/weekend	9.82	9.82	9.82	9.82

(b) WearGuard. WearGuard is a privately-held direct marketer and retailer of work clothing and uniforms with total sales in excess of $130 million per year. In addition to a string of wholesale and retail outlets, the firm publishes a catalog and accepts orders through its Telephone Sales Department which is located at the corporate headquarters in Norwell, Massachusetts. These phone lines are open twenty-four hours per day, seven days a week and are responsible for roughly one half of the company's catalog sales. The company produces some thirteen major catalog mailings each year. Since a majority of orders are placed within three weeks of a catalog's mailing, the mailings are staggered so as to maintain a steady arrival rate of calls and orders from day to day throughout the year.

Although the arrival rate of calls is fairly constant from day to day, there is a great fluctuation in calls per hour over the course of a day. Between midnight and 8:00 am, calls come in at a very low rate. During the business day, the rate steadily increases as establishments across the country open for business and then decreases through the evening hours as companies start to close. The distribution of call origins also varies throughout the day with more calls coming from the East in the morning and from the West in the evening. (Unfortunately, at the time of the study data were not available on call origins by time of day from the company, so a constant distribution is assumed; however, the modeling techniques that we develop can accommodate a non-homogeneous pattern of call arrivals.)

As do many telemarketers, the company makes use of AT&T's WATS line "hunt" system, which accepts calls and assigns them to the least expensive line which is free. For example, if a call were to arrive from Zone III, the system would try to assign the call to a Band III line. If all Band III lines were busy, the system would try to assign the call to a Band IV line, etc. If there were no Band III, IV, or V lines available, the caller would receive a busy signal. If a line is free but there are no available operators to serve the call, the system plays a recorded message and puts the caller on hold. It should be noted that during this time, the WATS line is in use and charges accumulate. The cost of the call reflects the band actually utilized, not the zone of call origin. In addition to assigning calls to lines, the system also maintains statistics of origins and lengths of calls as well as the number of calls that are abandoned and the length of time each caller spends "on hold" waiting for an operator to answer. By using the statistics provided by the

system, it is possible to generate most of the data required by the model we developed here.

(c) Problem definition. As mentioned in the last section, the hunt system assigns each incoming call to the least expensive available line which can service a call coming from a given origin. Clearly the total telephone costs depend not only on the number of phone lines, but also the configuration of phone lines for which the company contracts. This paper develops a model which can be used to determine the optimal configuration of incoming WATS lines to minimize telephone costs while maintaining an acceptable level of service. In the case of WearGuard, "acceptable level of service" is defined as a level at which no more than X% of all callers receive a busy signal during any given period of the day, where the service percentage X is a managerial decision variable.

To model this problem, we effectively ignore consideration of operators. Clearly, the number of operators and their scheduling influence the optimal configuration of lines. If the number of operators is reduced, callers spend more time on hold and more lines are needed. Similarly, an increase in staffing level makes it possible to deliver service of the same quality with fewer lines. A complete optimization problem would therefore solve the optimal staffing level as well as the optimal configuration of lines. However, we do not address the staffing question here. Rather, we assume that the staffing and scheduling of operators are secondary issues to the question of the number of lines; we assume that having decided how many lines to have, the operators are scheduled so that an acceptable level of service, measured in terms of expected time on hold, is maintained throughout the day. Furthermore, the service target for expected time on hold will be set to be very small (i.e., a few seconds), so that the effect of the scheduling of operators on the total time to service a call is small.

We have found very few papers on this subject. Morrison [4] has done some work that focuses on characterizing the queueing behavior, but he has not addressed optimization issues. Recently, we have come across a study performed at Cornell University by Lampbell [1], which is very similar to ours. Lampbell proposes the same queueing model as we do, for approximating the load on a given line configuration. He also develops optimizing and heuristic methods for selecting the number of lines. These methods differ somewhat from the dynamic program developed in this

paper. The methods given by Lampbell permit a more general relationship between the different types of WATS lines, in which the service areas of different line types need not overlap. Our dynamic program is specialized to the case faced by telemarketers using WATS lines, where the service areas are nested. One reason for the difference in assumptions is that Lampbell is concerned with lines for outgoing calls, whereas the WearGuard application is for incoming calls. Nevertheless, there is great similarity between our work and that of Lampbell.

The remainder of the paper consists of three sections. In section 2 we develop a queueing model to describe the system and then use dynamic programming techniques to find the best solution. Then in section 3 we describe the data required by the model, give an example from WearGuard, and check the sensitivity of the solution to various parameters of the model. Finally, in section 4 we validate the solution, first with an approximation to Larson's hypercube model, [2], [3] and then with a simulator which has been used at WearGuard to choose staffing levels.

2. MODEL DEVELOPMENT

The model development entails two components. First, we propose a queueing model for finding the usage levels for a given configuration of lines. We then embed this queueing model within a dynamic program, which determines the least-cost configuration of lines. We begin by presenting the queueing model.

(a) The queueing model. The first step in solving the problem is the development of a model that describes the system while maintaining some degree of tractability. In order to do so, it is necessary to make several simplifying assumptions. The first of these is to break the day into "time blocks" within which calls arrive as a Poisson process with a constant rate. Different time blocks have different arrival rates.

The second assumption states that the system is generally in a steady state. Within each time block, calls arrive at a constant rate and we assume that the system is ergodic in each time block. In moving from one time block to the next, the system may experience transient behavior before achieving a new steady state. As long as there are relatively few time blocks, each one of these blocks will be sufficiently long that the system will spend most of the

time in steady state. In building our model, we therefore assume that the time the system spends not in a steady state is small and can be ignored.

The third assumption involves the overflow of calls from one band to another. When all lines of a given band are busy, a call that arrives to that band overflows to the next. Morrison [4] has developed polynomial expressions to describe overflow queues in the two band case, and his work could conceivably be extended to describe a five band system, but the resulting equations would be very complex and optimization of the number of lines would seem impossible. To preserve tractability, we model the overflow of calls as an independent Poisson process. This is an approximation since actually the overflow process is a "censored" Poisson process: there is Poisson overflow to Band j only when Band j-1 is full. Since we model the overflow of calls from Band j-1 to Band j as Poisson and since calls arrive from Zone j in a Poisson manner, it follows that the arrival of calls to each band is the sum of two independent Poisson processes and is therefore itself Poisson. Lampbell [1] made the same assumption to approximate the overflow process as a Poisson process.

Our fourth assumption is that service times are independent and identically distributed. This assumption is valid if there is one operator for each phone line and hence, calls are never put on hold. If, however, there are fewer operators than lines, then callers will occasionally be put on hold for some period of time. This time on hold is dependent upon the arrival and duration of other calls, and thus, the service times would not be independent and identically distributed. Nevertheless, when deciding the number of lines, for service reasons we assumed that WearGuard would always staff enough operators either to cover all of the lines or at least to ensure that the time on hold is a very small portion of the total time spent being served.

One final assumption is that it is valid to neglect the fringe effects created by the price breaks. We will discuss this assumption later in the section. Using these five assumptions, we build the queueing model of the telephone system.

In constructing this model, we effectively assume that the number of operators available is equal to the number of lines. We model each band of lines as an $M/G/n_j$ queue with no waiting room, where n_j equals the number of Band j lines. Assume that calls arrive at a rate $\gamma_{jt} = \lambda_{jt} + \zeta_{j-1,t}$ in time block t where λ_{jt} is the arrival rate of calls from Zone j and $\zeta_{j-1,t}$ is

the overflow rate from Band j-1. Since we assume the overflow process is Poisson and we ignore the transient behavior, it follows that P_{jnt}, the probability of having n Band j lines busy at any time during period t is equal to

$$(1) \qquad P_{jnt} = \frac{(\gamma_{jt}/\mu)^n/n!}{\sum_{i=0}^{n_j} (\gamma_{it}/\mu)^i/i!} \quad ,$$

where $1/\mu$ is the average length of a call (e.g., Tijms [6] p. 361) and $0 \leq n \leq n_j$.

Given that this is true, calls will overflow from Zone j at a rate equal to the arrival rate of calls multiplied by the probability of having all n_j lines busy, i.e. Erlang's loss formula. That is,

$$(2) \qquad \zeta_{jt} = \gamma_{jt} P_{jn_jt}$$

where P_{jn_jt} is determined by (1) with $n = n_j$. Hence, we can use ζ_{jt} to determine the total flow to Band j+1, etc.

To calculate the expected cost of a given configuration, we need to determine the expected monthly usage cost. To compute the expected usage fee, we need to know the total monthly usage of each band in order to charge the proper rate. Yet, the total monthly band usage is a random variable so the rate is also a random variable. As a simplification, we calculate the expected monthly usage fee by assuming the usage rate that corresponds to the expected usage level. Such an approximation is very easy to calculate and gives a good estimate of the expected cost except for the case in which E(HU) is very close to a price break and both Pr(HU < price break) and Pr(HU > price break) are significantly greater than zero, HU being the monthly usage level. Even in such a case, the error is less than the savings associated with the price break which is generally equal to about 10% and is never greater than 13% of the usage cost. Due to other approximations in the model, an error which must be significantly less than 10% seems tolerable.

The expected number of lines that are busy for any time block and band is

(3) $$\left(\sum_{i=1}^{n_j} iP_{jit}\right) .$$

The monthly usage for a time block is simply the product of the expected number of busy lines and the number of hours in that time block over a month. The total monthly usage is the sum of monthly usages for all time blocks. We then use the total monthly usage, along with the rates from Table 1, to determine the appropriate WATS charge. Using this model, we generate a good approximation of the cost of the phone system as a whole.

To determine the service level provided by a given line configuration, we use the overflow rate for Band V, as given by (2) for j = 5. This overflow rate is the number of calls per hour that are blocked or lost (i.e., receive busy signal) in time block t.

(b) The dynamic program. The problem is to determine the optimal number of lines for each of five telephone bands so as to minimize the total cost of the system. Such a problem lends itself well to a dynamic programming approach in which an optimization problem is divided into a series of stages, each of which is solved sequentially to arrive at a final solution.

We define the system cost function $f_{im}(n)$ as the total cost for Bands 1,2,...,i with a total of m lines where Band i has n lines ($n \leq m$) and the remaining m-n lines are optimally assigned to bands 1,2,....,i-1. Given this definition of $f_{im}(n)$, we define f_{im} as $\min_n (f_{im}(n))$, and n_{im} as the value of n that minimizes $f_{im}(n)$. Thus, f_{im} is the total cost for bands 1,2,....i with the optimal assignment of m lines.

There are four components to the cost function $f_{im}(n)$: the access fee for the n Band i lines, the usage cost for Band i, the overflow cost for serving calls which arrive to Band i but overflow to a higher band, and the cost associated with the first i-1 bands.

The access fee for n Band i lines is the fixed monthly cost for having n lines, which we denote by $AF_i(n)$. This cost is linear with respect to n and is the same for all bands.

The usage cost for Band i depends on the call arrival rate to Band i, and on the number of Band i lines. The call arrival rate to Band i is the sum of the Zone i arrival rate and the overflow rate from Band i-1, which depends on the lines configuration for Bands 1,2,...,i-1. We let $\zeta_{it}(m)$ denote the overflow rate from Band i in time block t, given that m lines are optimally allocated to Bands 1,2,...,i. Then, the arrival rate in time period t to Band i, given that m-n lines are allocated optimally to Bands 1,2,...i-1, is

$$\gamma_{it}(m-n) = \lambda_{it} + \zeta_{i-1,t}(m-n) \quad . \tag{4}$$

We then use these arrival rates in (1) to determine the usage rates in each time block for the n Band i lines, from which we estimate the usage cost as described previously. We denote by $UC_i(m,n)$ the usage cost for Band i with n lines, and m-n lines allocated optimally to Bands 1,2,....,i-1.

We need to include in $f_{im}(n)$ the cost for the calls which arrive to Band i, but overflow to a higher band. To estimate this overflow cost, we assume that calls which overflow are served by the next higher band and are charged a usage fee that is prespecified. Typically, we assume that the usage fee corresponds to the rate charged for usage of eighty hours per line-month. We note, though, that this is another approximation since we cannot know how the overflow calls will be handled until we configure the higher bands. We denote our approximation to the overflow cost by $OF_i(m,n)$ for Band i with n lines, and m-n lines allocated optimally to Bands 1,2,...i-1.

We need to note one exception to this treatment of overflows. An overflow from Band V is not handled by a higher band, but is a busy signal for the caller. Hence, there is not a direct charge associated with an overflow from Band V, but there may be a very large indirect cost due to a lost call. The customer may not call back, in which case the cost for an "overflow" is the cost of a lost sale. Furthermore, due to the poor service the customer may be less likely to try to call with subsequent orders, so future sales may also be lost. Rather than try to estimate these costs, WearGuard sets a service target of X%; that is, during any time period no more than X% of the arriving calls should receive a busy signal. This approach is typical for telemarketing operations. This constraint is easy to incorporate into the evaluation of $f_{im}(n)$ for Band V (i=5): if the overflow rate exceeds X%, we set $f_{im}(n) = +\infty$

The final component of the cost function is the cost associated with the first i-1 bands. For the dynamic programming recursion, we assume that we have previously determined the best allocation of the m-n lines over Bands 1,2,...,i-1 and that we know the expected monthly cost for this allocation, given by $f_{i-1,m-n}$. However, this cost includes the estimate of the cost for handling the overflow from Band i-1 by the higher bands. For computing the costs through Band i, $f_{im}(n)$, we need to subtract the estimate of the overflow cost used in finding $f_{i-1,m-n}$ to avoid double counting. Thus, we denote by $g_{i-1,m-n}$ the expected monthly cost for Bands 1,2,...,i-1 with the optimal allocation of m-n lines, exclusive of the costs for calls from Zones 1,2,....,i-1 that overflow from Band i-1 to Band i or higher.

Now we can write the recursion for $f_{im}(n)$, for $0 \leq n \leq m$:

$$f_{im}(n) = AF_i(n) + UC_i(m,n) + OF_i(m,n) + g_{i-1,m-n}, \tag{5}$$

$$f_{im} = \min_{n} \{f_{im}(n)\} = f_{im}(n_{im}), \tag{6}$$

and

$$g_{im} = f_{im} - OF_i(m, n_{im}). \tag{7}$$

For Band V, an overflow results in a lost call. In this case, we constrain the overflow rate to be no more than X% of the total calls that arrive to the system during any time block.

Given these formulae, we find the optimal solution by solving the equations for Band I and working out to Band V. For each successive value of i, we solve equation (5) for m=0,1,...., $\overline{m}$ where $\overline{m}$ is such that the overflow rate of calls is less than an arbitrarily small value (in the case of this study, a value of .01 calls/hour is employed). The minimum value of f_{5m} over the various values of m represents the optimal solution for the system. Let m^* be defined as the optimal total number of lines ($f_{5m^*} \leq f_{5m}$ for all m). The optimal number of Band V lines would then be given by $n_5^* = n_{5m^*}$. The optimal number of Band i lines is equal to

(8) $$n_i^* = n_{ik}, \text{ where } k = m^* - n_5^* - \ldots - n_{i+1}^*$$

Using this method, we estimate the total expected monthly cost ($=f_{5m^*}$) and find the optimal number of lines for each band.

This dynamic programming algorithm allows for a fairly efficient method of solving the optimization problem. Assume that M is the maximum conceivable value of m. For each value of m, a maximum of m+1 functions must be evaluated as n ranges from 0 to m. Therefore, we can anticipate a total of $\sum_{i=0}^{M} i+1$ evaluations for each band, which means the algorithm will perform as a function of $O(KTM^2)$ where K is the number of bands in the system and T is the number of time blocks. This is far more efficient than complete enumeration.

We have implemented the algorithm in FORTRAN, and the complete code appears in Santos [5]. This program requires as input the number of bands and time periods in a day, the average service time for a call, the access and usage fees, the maximum allowable loss rate, the arrival rates of calls from each zone during each time block, the number of hours of each time block in a month, and the presumed usage rate for each band and time block. The program reads this data from an input file, solves the problem using the algorithm described in this section, and gives as a solution the optimal number of lines in each band, the overflow rate from each band during each time block, and the total cost of the system (excluding overflow cost) up to and including that band. A typical problem requires less than one minute of CPU time to be solved on a PRIME 850 computer.

3. EXAMPLE OF MODEL APPLICATION

In this section we present and discuss the initial example that we worked with when building and testing the model in 1984. For this example we first describe the input data required by the model, and then present the optimal solution from the dynamic program. We compare this solution with an alternate solution, namely the line configuration which WearGuard had at

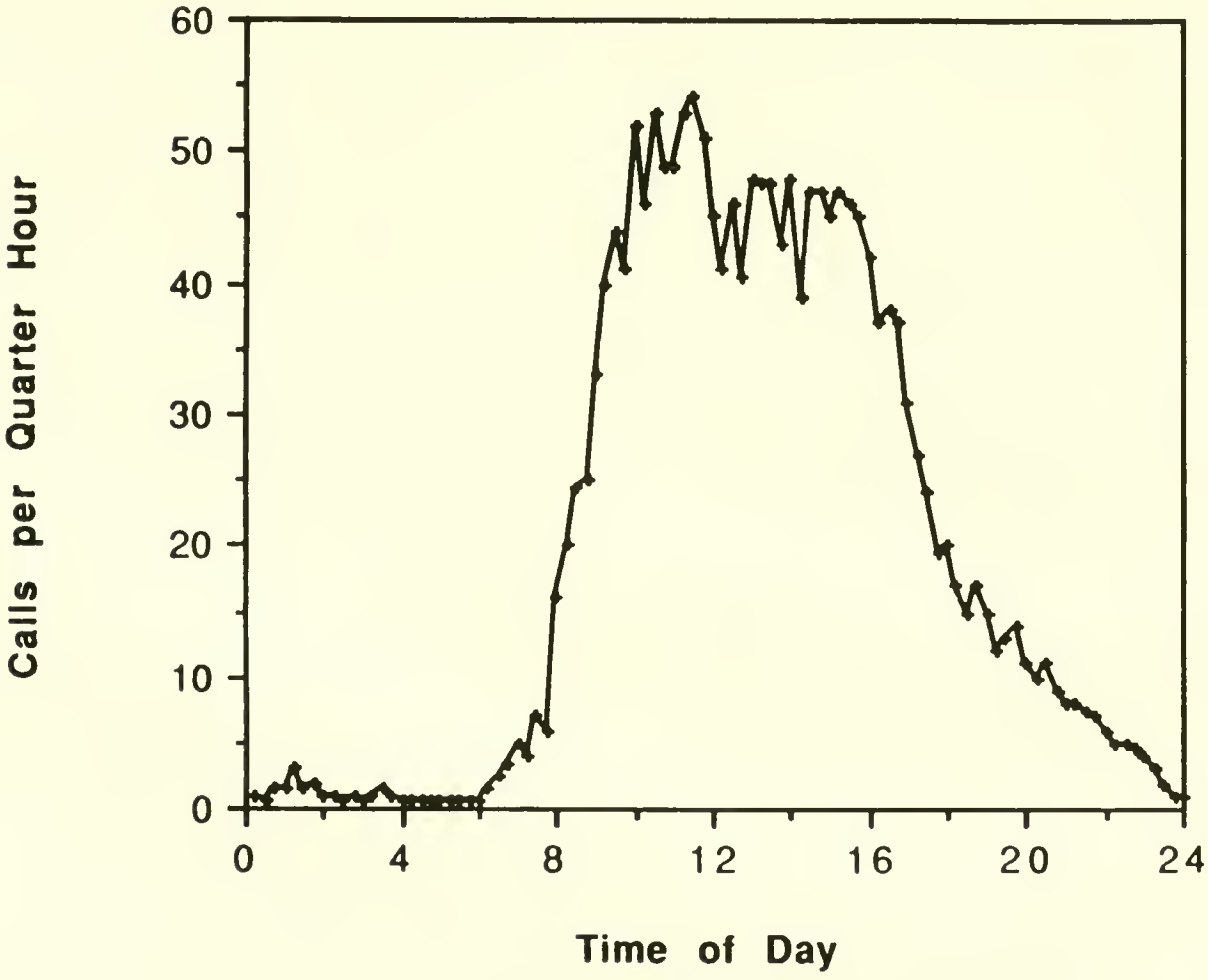
the time of the study. Finally, we examine the sensitivity of the solution to various changes in the input data.

(a) Data. For the given model, we can solve the configuration problem for any operation for which the necessary input data are available. The access fees and usage rates are set by the telephone company; we provide an illustration of these in Section 1.1 and Table 1. Other data, namely arrival rates, time blocks, source distribution, and service time information are specific to the firm being studied and must be extracted from their available data sources. The final input parameters are the estimated overflow penalty and the maximum percentage of calls lost. We initially set 5% as an acceptable rate of lost calls during the peak period. Thus, any solution with a maximum or peak loss rate greater than 5% is considered to be not feasible. The estimated overflow penalty is set equal to the cost of serving a call at the next higher band at the maximum price break (i.e., more than eighty hours per month).

The parameters that remain are specific to the company in question. As was mentioned previously, the telephone system used by WearGuard maintains relevant statistics which are tabulated every day. These statistics are maintained on Daily Summary Sheets which contain such data as the average length of a call and the average time spent on hold, and Daily Profile Reports which contain the arrival rate of calls for each fifteen minute interval throughout the day. Using these summary sheets over the course of several weeks, along with other data which has been collected in studies performed previously at the company, we determined the input parameters. We describe next the values for arrival rates, distribution of call origins, and mean service times.

The arrival rate data define how many calls are received over the course of the day and when during the day they are received. As an illustration, a graph of these arrival rates appears in Figure 2, which gives the arrival rate of calls for each fifteen minute interval over a day. There is remarkably little variation in the arrival rate pattern from day to day during the week, and the distribution has proven to remain stable over the past five years (although daily call volumes have increased). We considered the cost of operating the telephone system over the weekend negligible and therefore ignored it. We instead assumed that an average month consisted of close to twenty-two days, each with the same arrival rate of calls.

Figure 2: Arrival Rate of Calls



From this data we divided the day into time blocks within each of which we assumed a constant arrival rate of calls. We needed a sufficient number of blocks to describe the changing rates throughout the day while having few enough blocks to keep the problem tractable. Natural breaks occur at 7:00 am, 5:00 pm, and 11:00 pm when usage fees change. Since the absolute peak period is a major source of expense and represents the time when the greatest number of calls will be lost, it was necessary to isolate that period into a unique time block. We therefore broke the business day (7 am to 5 pm) into two blocks, one consisting of the peak period and running from 10:15 am to 12:00 noon, and the other consisting of the rest of the business day (i.e., 7:00-10:15 am and 12:00-5:00 pm). During the evening hours, the arrival rate of calls decreases monotonically from 40.9 calls per fifteen minutes at 5:00 pm to 3 calls per fifteen minutes at 11:00 pm. It seemed prudent to split the evening into two time blocks of equal length, one running from 5:00 pm until 8:00 pm and the other from 8:00 pm until 11:00 pm. The arrival rate between 11:00 pm and 7:00 am is low and generally fairly constant. For this reason, only one time block was used for that interval. The arrival rates for an entire day were thus compacted into five time blocks, the data for which appear in Table 2.

The data for this sample period indicated an arrival rate of 2283 calls per day, distributed as displayed in Figure 2. However, at the time of the study, WearGuard desired to configure the phone system to accommodate thirty-five hundred calls per day, which represented future call volumes. (Call volumes in 1989 are around 5500 per day.) Thus, for an arrival rate of 3500 calls per day, we adjusted each of the observed arrival rates. These arrival rates also appear in Table 2 in the column titled "Adjusted Arrival Rate."

Table 2: Time Block Data

Time Block	Time	Hours/ Day	Hours/ Month*	Calls/ Hour (2283 per day)	Adjusted Arrival Rate (3500 per day)
1	0.00-7:00; 23:00-24:00	8	173.9	6.0	9.2
2	10:15-12:00	1.75	38.0	215.8	330.8
3	7:00-10:15; 12:00-17:00	8.25	179.4	186.9	286.4
4	17:00-20:00	3	65.2	73.8	113.2
5	20:00-23:00	3	65.2	31.5	48.3

*We assumed 21.74 days/month

After defining how many calls come into the system, the next step was to determine call origins. The distribution of sources is clearly not constant over the day. In the morning, when the arrival rate first starts to rise, calls are predominantly from the East as most businesses in the western part of the country have yet to open. Similarly, in the evening hours as the overall call rate starts to decline, a larger percentage of the calls come from the West. Although the dynamic program takes as input the number of calls per hour from each zone and could easily handle a changing distribution over the span of a day, WearGuard did not have data broken down in such a way as to make a distinction. Therefore, for the purpose of this example, we assumed a constant distribution of origins over the course of the day. This distribution appears in Table 3.

Table 3: Distribution of Call Origins

Zone	Percentage
I	21.7%
II	21.0%
III	23.7%
IV	12.2%
V	21.4%

The final piece of input data is the service time, or average length of call. This includes both the time spent on hold waiting for an operator to answer the call and the service time once the call has been answered. For WearGuard this number was determined from their Daily Summary Sheet where average call times were tabulated each day. A weighted average of average call times was obtained from these summary sheets from fifteen weekdays between December 1, 1983 and January 6, 1984. The average length was 3.83 minutes, or 15.2 calls per hour, based on a sample of 31,614 calls.

(b) Results. Using the data presented in the previous section, we determined the optimal configuration of phone lines for WearGuard. The results of the run for the base case appear in Table 4 where we see that the optimal configuration consists of 12 Band I lines, 12 Band II lines, 14 Band III lines, 7 Band IV and 5 lines for Band V. This 12-12-14-7-5 solution represents an entirely different strategy from the 2-9-2-10-20 configuration which was in place at the time (see Table 5; we used the queueing model described earlier to compute the costs and the overflow rates given in Table 5 for the exisiting configuration with a call rate of 3500 calls per day). In both Tables 4 and 5 we list for each band the peak overflow rate, which is the largest overflow rate over the time blocks in a day for the band and is computed from equation (2). The existing configuration employed the strategy of using several Band V lines to catch overflows while employing relatively few lines on the lower bands to save on access fees. The optimal strategy, however, has many lines at the lower bands to keep calls from overflowing to the more expensive outer bands and leaves relatively few lines on the outermost bands. Band V

serves primarily calls from Zone V and most lost calls come from Zones IV and V. Such a strategy results in more lost calls (as close to the constraint as is possible) and requires more lines, but seems to save money.

The optimal solution has a peak overflow (lost-call) rate from Band V of 16.3 calls per hour; however, this peak rate holds for only two hours of each day. Table 6 shows the rate of lost calls in the other time periods of the day. Over the remainder of the business day, roughly 3.7% of all calls are being lost. This overall loss rate is well within the constraint of 5%. However, the peak loss rate of 16.3 calls per hour was considered high enough to warrant lowering this 5% constraint, which we examine in the next section.

Table 4: Results for Base Case

Band	Number	Access Fee	User Cost	Total Cost Cumulative	Peak Overflow Rate Calls/Hour
I	12	442	12,930	13,372	0.2
II	12	442	13,837	27,651	0.1
III	14	515	15,987	44,154	0.0
IV	7	258	8,311	52,723	0.4
V	5	184	12,074	65,003	16.3

Table 5: Costs of Existing Configuration

Band	Number	Access Fee	User Cost	Total Cost Cumulative	Peak Overflow Rate Calls/Hour
I	2	74	5,616	5,690	49.4
II	9	331	19,437	25,458	19.5
III	2	74	6,448	31,980	68.5
IV	10	368	18,855	51,203	7.5
V	20	736	17,833	69,772	0.0

Table 6: Loss Rates of the Base Case Solution

Time Block	% of Daily Arrivals	Lost Calls/ Hour	% of Calls
1	2.1	0.0	0.00
2	16.5	16.3	4.93
3	67.3	10.5	3.67
4	9.7	0.4	0.35
5	4.1	0.0	0.00

(c) Sensitivity. In the previous section, we developed and solved an example based on the operating conditions at WearGuard in 1984. However, there was some ambiguity in the definition of some of the parameters of the problem. To check the robustness of the solution, we examine the sensitivity of the solution to variations in the service constraint, arrival rate of calls, and average length of call. Table 7 shows the optimal solution to the problem under a variety of scenarios in each of which one of these parameters was shifted away from the value defined in the previous section.

We first considered the sensitivity of the results to the service overflow constraint. The optimal solution with a maximum overflow rate of 5% allows for a loss rate of 16.3 calls per hour during the peak period, and 10.5 calls per hour over the rest of the business day, for a loss rate of roughly 115 calls per day. Such a high loss rate not only impedes service for these 115 calls, but as those callers who received busy signals call back, the arrival rate of calls will increase, possibly causing an even greater loss rate. For these reasons, we wanted a tighter constraint. We reran the problem with the service constraint set at 3%, 1%, and 1/2%. From the results in Table 7, we see that as the constraint becomes tighter, the optimal solution is attained by adding lines to the outer bands to reduce the loss rate while maintaining the original structure of the solution for the inner bands. The maximum overflow rate can be reduced from 5% to 1% for an increased cost of $2100 per month, or 3.2% of the total system cost, or down to 1/2% if total system costs

TABLE 7: SENSITIVITY ANALYSIS

CALLS/DAYS	3500	3500	3500	3500	3850	3150	3500	3500	5500
Service Time	3.83	3.83	3.83	3.83	3.83	3.83	4.21	3.45	3.83
Loss Rate	.05	.03	.01	.005	.01	.01	.01	.01	.01
Band I	12	12	12	12	14	11	14	11	19
Band II	12	12	12	12	14	11	14	11	18
Band III	14	8	14	14	15	12	15	13	19
Band IV	7	0	6	3	5	4	5	4	9
Band V	5	10	8	10	9	8	9	8	11
Total Lines	50	42	52	51	57	46	57	47	76
Peak Over-flow Rate	16.3	9.6	3.3	1.6	3.1	2.5	2.9	2.9	5.0
Cost ($K)	65.0	66.2	67.1	67.3	73.8	60.4	74.3	61.3	105.3
% Overflow at Peak	4.9	2.9	1.0	.5	.9	.8	.9	.9	1.0

are increased 3.5%. After some discussion, it was decided that a solution with a lower overflow rate was preferable to the 5% rate originally assumed. For subsequent sensitivity analysis, we assumed a service constraint of 1%.

We next considered the effect from variations in the arrival rate. Additional runs were performed with arrival rates of 3850 (10% greater than the base rate) and 3150 (10% less) and 5500 calls per day. The results of these runs also appear in Table 7. When the arrival rate increases by 10%, seven lines are added to the system, two each on Bands I and II and one on Band III, so as to maintain a very low level of overflow to the outer bands. Two lines are also added to Band V and two are shifted from IV to V so as to keep the loss rate below the constrained value of 1%. Such a system would have an expected cost of $73,800 per month which represents an increase of 13.5% over the base case with a maximum overflow rate of 5%; the increase in expected cost is only 10% over the base case with a 1% maximum overflow rate.

When the arrival rate is reduced to 3150 calls per day, the opposite occurs. One line is removed from each of the first two bands, two lines are removed from Band III, but three lines are moved from Band IV to Band V. The resulting 11-11-12-4-8 configuration has an expected cost of $60,400 per month which is about 7% less than the cost of the base system with a 5% maximum overflow rate; the expected cost is 10% less than the base case with a 1% maximum overflow rate.

Lastly, we considered the sensitivity of the solution to changes in the service time. The service time for this problem includes both the time required to serve the customer and the time the caller spends on hold waiting to be served. In developing the model, we assumed that the time on hold constitutes a very minor portion of the total service time. Over the fifteen days for which data were tabulated, the average caller spent four seconds on hold before being served. Nevertheless, a change in operator scheduling could have an effect on the service time due to this delay time. For this reason additional runs were made to observe the sensitivity of the solution when the service time was increased or decreased by 10%. From the results in Table 7, we see that a 10% increase or decrease in service time has a very similar effect as a 10% shift in arrival rate. The resulting strategy is very close or the same, and the monthly costs vary by less than 1/2%.

4. MODEL ACCURACY

In this section we examine the accuracy of the model. This will be done by comparing the estimated cost from the model to that from another approximate model and from a simulator that is used at WearGuard for the purpose of scheduling operators.

(a) Approximation of the hypercube model. The hypercube model is an analytic model, developed by Larson [2], [3] for the purpose of analyzing spatially distributed queues in emergency vehicle systems. The model works by setting up a priority schedule for each zone of a city indicating the order in which emergency vehicles should be dispatched given that the higher priority vehicles have been previously dispatched to other emergencies. This system is completely analogous to the telephone system that has been described in this paper where the different telephone bands correspond to zones of a city and the telephone lines are similar to emergency vehicles. The hypercube model makes several assumptions. Those that remain relevant in the conversion from emergency vehicles to telephone lines are the following:

(1) Independent Poisson arrivals.
(2) N servers, each of which can travel to any zone in the region.
(3) Single server dispatch to any call; a call is lost if all servers are busy.
(4) Fixed preference dispatching.
(5) Exponential or near exponential service times.

Of these assumptions, the only one that is contradictory to the telephone system is the second one. This claims that each of the telephone lines could handle any call while in fact a call from Zone j can only be served by Bands j, j+1,....,5. This problem can be avoided by creating a buffer of "Band VI" phone lines which would be placed in the priority lists for calls arriving from Zones II through V after the Band V lines but before the Band I lines. For calls from Zone I, the Band VI lines would be placed at the very end of the priority list. A call can only be served by an infeasible line if all the Band V lines are busy and all the buffer lines are busy as well. If enough buffer lines are included, the probability of an overflow to a Band I line becomes insignificant. The assumption of exponential service times is not needed by our queueing model, whereas it is needed for the hypercube model. However, the actual

service times are probably close enough to being exponential to make this assumption reasonable.

The hypercube model describes the system exactly, but requires that 2^N simultaneous equations to be solved. For this reason, the problem becomes intractable for large values of N, e.g., N greater than fifteen servers. Larson has developed an approximation to the model which requires that only N equations be solved simultaneously and which generally solves the problem to within one or two percent of the exact results. Since the problem being considered has a value of N equal to about 50, we use this approximate solution procedure. Table 8 compares the results of the hypercube evaluation for the 12-12-14-7-5 configuration to that for our approximate queueing model. The model uses an iterative

Table 8: Comparison of User Cost Estimates Using the Hypercube Model and the Approximate Queueing Model

	Estimated User Cost		
Band	Hypercube Model	Approximate Queueing Model	% Deviation
I	13,252	12,930	-2.5
II	13,791	13,837	.3
III	15,727	15,987	1.6
IV	8,224	8,311	1.0
V	13,090	12,074	-8.4
Total	64,084	63,139	-1.5

procedure to solve the problem, which typically requires few iterations to converge. Unfortunately, in the example studied here, the solution failed to converge for arrival rates corresponding to the peak time block. The solution for that time block, therefore, corresponds to that which was attained on the iteration that had a solution that varied the least from that of the previous iteration. There is no way of knowing how good the hypercube

approximation is. However, since the estimated costs of the two models for each of the first four bands vary by less than 2.5% and the expected user cost of the entire system as calculated using the hypercube model is only 1.5% greater than the expected cost from the approximate queueing model, the two models do not contradict each other and in fact there is a very high level of consistency.

(b) Simulation model. In addition to the hypercube model, another method of testing the accuracy of the approximate queueing model is through simulation. A time-driven simulator was developed by WearGuard for the purpose of scheduling operators. Instead of dividing the day into large time blocks during which the arrival rate of calls remains constant, the simulator has a different arrival rate for each fifteen-minute time block over the day. A second difference between this model and the one developed in this paper is the fact that the simulator considers the service time to be a normally distributed random variable (truncated at zero) with mean and variance equal to 3.83 minutes. The simulator takes as input the configuration of lines, the number of operators over the course of the day, and the duration of time to simulate. By choosing the number of operators to be equal to the number of telephone lines for a one-month simulation, we obtain from the simulator an estimate of the monthly cost of a given configuration. One advantage of checking our solutions with the simulator was that the simulator gave a very accurate calculation of the costs.

We used the simulator to estimate the cost of both the optimal solution and the current system. We also used the simulator to determine whether or not an improved solution could be found by adding or removing one line from each of the bands. The results of this procedure are in Table 9. Once again the estimated cost using this simulator comes very close to the estimate of the queueing model; they differ by less than 2%. Since the result of the simulator, the hypercube model, and the approximate queueing model all fell within a range of 3% of each other, we concluded that the queueing model developed in this paper gives an accurate result.

Table 9: Costs of Various Configurations as Estimated Using the WearGuard Simulator

	Configuration Bands (I-II-III-IV-V)	Access Fee($)	Usage Fee($)	Total Cost($)
Optimal	12-12-14-7-5	1840	62,045	63,885
Add Band I	13-12-14-7-5	1879	63,521	65,400
Drop Band I	11-12-14-7-5*	1803	61,320	63,123
Add Band II	12-13-14-7-5	1879	63,640	65,519
Drop Band II	12-11-14-7-5*	1803	63,067	63,870
Add Band III	12-12-15-7-5	1879	61,975	63,854
Drop Band III	12-12-13-7-5*	1803	59,381	61,184
Add Band IV	12-12-14-8-5	1879	62,188	64,067
Drop Band IV	12-12-14-6-5*	1803	60,953	62,756
Add Band V	12-12-14-7-6	1879	63,205	65,084
Drop Band V	12-12-14-7-4*	1803	60,641	62,444
Current	2-9-2-10-20	1582	64,902	66,484

*Indicates a peak loss rate greater than 5%

The final application of the simulator was to test if the solution found with the optimization model actually represents a local optimum. From Table 9, we see that removing a line reduced cost but causes the loss rate to exceed the constraint of 5%. When we add a line, the expected costs increase, with one exception. The line configuration, 12-12-15-7-5, costs sightly less than that for the configuration found by the dynamic program; however, the percentage cost difference is less than 0.05%. We therefore concluded that the optimization model developed in this paper finds a solution which is either optimal, or very close to optimal.

References

1. Lampbell, David M., "On the Selection of Numbers of Servers for the N Server-Type Problem," Technical Report No. 349, School of Operations Research and Industrial Engineering, Cornell University, Ithaca NY, August 1977.

2. Larson, Richard C., "A Hypercube Queuing Model for Facility Location and Redistricting in Urban Emergency Services," Computers and Operations Research, 1 (1) 1974, pp 67-95.

3. Larson, Richard C., "Approximating Performance of Urban Emergency Service Systems," Operations Research, 23 (5) September-October1975, pp 845-868.

4. Morrison, J. A., "Analysis of Some Overflow Problems with Queuing," The Bell System Technical Journal, 59 (8) October 1980, pp 1430-1434.

5. Santos, P. Clark, "An Optimization Model for the Configuration of Incoming WATS Lines," unpublished M.S. Thesis, A. P. Sloan School of Management, Massachusetts Institute of Technology, Cambridge MA, 1984.

6. Tijms, Henk C., Stochastic Modelling and Analysis: A Computational Approach, John Wiley & Sons, Chichester, Great Britain, 1986.