W. H. Evers
S. S. Thakur

Programmed Automatic Customer Engineer (PACE) Dispatch

Abstract: An experimental real-time system is described for assigning customer engineers (servicemen) to requests for service, preventive maintenance and engineering- and sales-change activities. The system, which can be applied to service organizations of many kinds, is viewed as a stochastic programming formulation. The resultant mathematical programming problem is structured as a control system, an inner control loop and an outer adaptive feedback loop in which system parameters are adjusted based on a performance index. Tests of the system have been made using data from the Brooklyn, New York and Washington, D. C. IBM Field Engineering Division branch offices.

Introduction

A problem essential to providing adequate service to customers is the efficient allocation (dispatching) of customer engineers (CE's) to service calls on a real-time basis. Other customer engineering functions which must be considered when allocating CE's are routine preventive maintenance and the addition of sales and engineering changes to existing machines.

Figure 1 shows the (highly simplified) basic operation of a typical dispatch center. A request for service from a customer enters the system; if the primary customer engineer assigned to that account is available, he is notified and dispatched. If he is not available, the dispatcher attempts to contact the nearest available qualified CE in the territory. If no CE is available in the territory, the territory manager is contacted and, if urgency warrants it, arrangements are made to obtain a CE from another territory. This cycle illustrates the dispatch operation in its simplest form.

The broadening data-processing product line is fostering increased product specialization on the part of customer engineers. As a result the territorial assignment of the customer engineer is changing from one which is oriented to few customers and many products to one with few products and many customers. This factor, coupled with the anticipated rapid growth in customer engineering demand, is greatly increasing both the complexity and the volume of the dispatching job.

Programmed automatic customer engineer (PACE) dispatch is an experimental real-time control system designed to allocate CE's to customer accounts such that customer service levels are maintained while the cost of operating the service organization is kept at a minimum. The allocation of CE's to customer requirements for service is viewed as a stochastic programming problem. The solution methods are based on concepts of adaptive control theory.

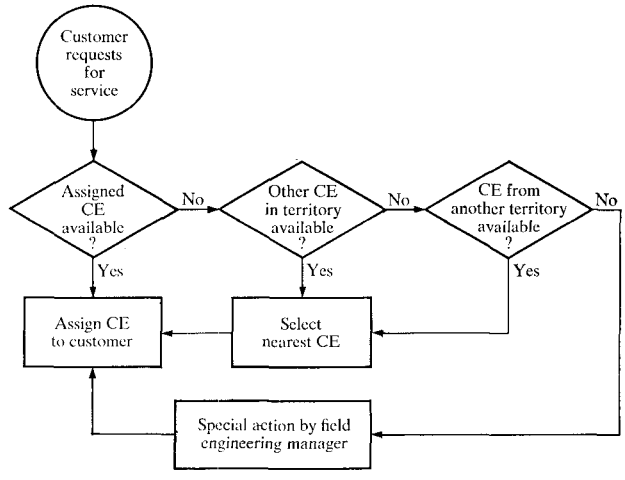

Figure 1 Customer-engineer dispatch-center operation.

PACE dispatch

The algorithms in this system are designed to provide service for each customer-machine or account-machine (AM) combination at prespecified levels. The seven operating factors of prime importance in specifying the

The authors are located at IBM Corporate Headquarters, Armonk, New York 10504.

Table 1 Principal operating factors affecting service level and cost of customer-engineer activities.

Factor	*Decision variable*	*Field engineering input*
CE system value (true travel cost)	Distance between AM's CE skill level	Addition and deletion of AM's Update CE skills
Overtime	Conditional expected overtime based on probability distributions of time to complete current job, travel time and next assignment	Update CE skills Maximum CE overtime Update CE shift status Update CE home-office assignments Update AM repair distribution
Response time	Expected time to complete present job based on CE estimate and probability distribution of completion times Customer priorities for service, i.e., response time for each AM Travel time to next AM	Required response time for each AM Maximum probability that response time is exceeded
Territory management	CE's distance when outside his territory CE's time outside his territory	Update CE territorial assignments Upper bounds on extra-territorial travel and time
AM-CE relations	Time since last CE visit Average number of CE visits per service call (including preventive maintenance and sales and engineering changes)	Average fraction of AM calls that primary CE should answer
CE capability	Type of machine to be repaired	Update CE skills
Preventive maintenance	AM to be serviced	Update territorial AM list Preventive maintenance schedule

service level and cost of CE activities are presented in Table 1. The PACE-dispatch-model cost of performing service at these levels is minimized by the system. To perform service for each AM, the program locates the "best" CE, defined as that CE with the necessary skills who can provide customer-acceptable service such that current or better-than-current service levels are maintained while the modeled cost of performing this service is minimized. The system allows field engineering management to maintain control by specifying the commitments the service organization will make to customers and service personnel. Of course, increasing the level of service increases the cost of the customer engineering operation.

Figure 2 illustrates in block form the operation of the algorithm. Customer requests for service enter in the form of a request arrival stream. At time t the system is said to be in state L_t. The system state is composed of the state of the CE, i.e., working or not, location of the node (AM) and the AM state, i.e., machine in need of repair or not. Based on the system state, a forecast matrix is generated. The element in row k and column j of the qth matrix is the predicted qth operating factor if CE k is assigned to AM j. The requests for service that arrive in the system enter a multichannel service system in which each channel represents a specific CE skill; a request queue may thus be expected to form. No ordering is imposed on the queue, but instead the loss or penalty associated with the estimates of the operating factors are calculated from loss or penalty functions. The deterministic cost $(c_{kj} \mid L_t)$ and the qth penalty cost $(p_{qkj} \mid L_t)$ of assigning CE k to AM j are obtained as functions of the state of the system L_t. The queuing problem can then be viewed as an assignment or allocation problem with associated cost matrix $(c_{kj} + \sum_q p_{qkj}|L_t)$.

The basic computational procedure for the *local* or *internal* control loop is to assign CE's to service requests based on these current system costs. A primal algorithm due to Balinski and Gomory[1] has been programmed for this system. This primal algorithm was chosen so that limits could be placed on computation time. For large matrices a near-optimal solution can thus be accepted if computation time exceeds the maximum allowable. In addition to the advantages of being a primal method, the algorithm provides a rapid computational method of solving assignment problems.

The *secondary* or *adaptive* loop is designed to obtain the values of $(p_{qkj}|L_t)$ that satisfy the constraints of the system with prespecified probability. The prespecified probability is viewed in this system as the performance criterion. The constraints are viewed as chance constraints and are satisfied probabilistically by choosing appropriate loss functions for the inner-loop allocation algorithm. The loss (penalty) associated with each operating factor is obtained from these loss functions. These functions are in turn functions of control parameters α_{qj}, which will be referred to simply as parameters when there is no

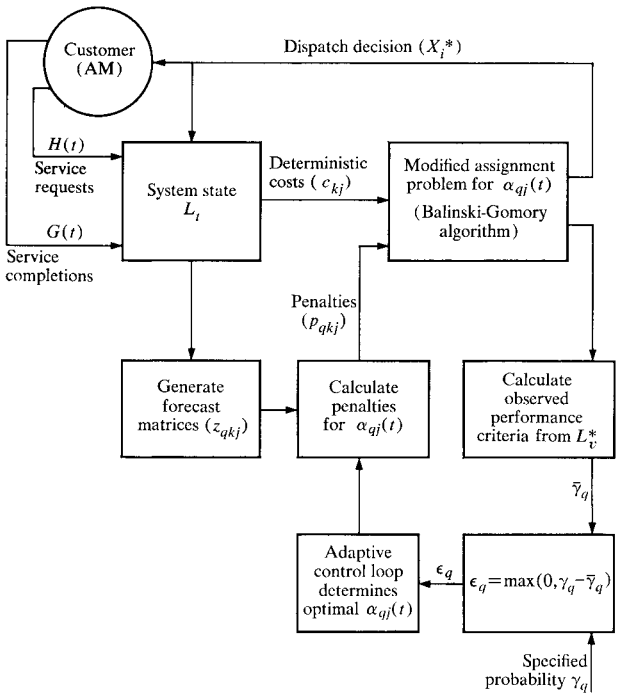

Figure 2 Flow diagram of the algorithm.

ambiguity. The adaptive loop continuously tracks the operating factors of the system and obtains those values of the control parameters that will satisfy the performance criteria. The terms performance criteria, control parameters, penalty functions, etc. are defined precisely later.

System description

Consider a typical service-repair facility consisting of N distinct nodes or account-machine (AM) combinations and a servicing crew of M CE's (repairmen). We are concerned with a system having finite-state space that will be controlled over a finite horizon. The system is observed at time t, $t = 0, 1, \cdots, w$, to be in state L_t. After observation, the system is "controlled" by making decision X_t, e.g., to assign $m_t \subset M$ to service $n_t \subset N$; m_t and n_t both have l elements, where $N \geq l$ and $M \geq l$; $X_0, X_1, \cdots, X_w$ is the sequence of decisions and $L_0, L_1, \cdots, L_w$ is the sequence of system states. Let

$$L_v^* = \{L_i : i = i - v, i - v + 1, \cdots, i - 1\}$$

be a subset of v states from the sequence of previously observed system states. Then, given that L_v^* has occurred, a decision X_i is made at $t = i$. The operating factors (in terms of CE $k \in m_i$ and AM $j \in n_i$ associated with X_i) can then be written as functions of the current decision X_i.

Customer engineers are normally grouped into subsets $N^1, N^2, \cdots, N^u$ such that T_u, the territory or set of AM's associated with N^u, has the following properties:

$$T_u \subset N, \quad \bigcup_u T_u = N \quad \text{and} \quad \bigcap_u T_u = \emptyset.$$

Further, subterritories are normally allocated to CE's (called prime CE's) such that subterritory T'_k associated with CE $k \in N^u$ has the properties

$$T'_k \subset T_u, \bigcup_k T'_k \subseteq T_u \quad \text{and} \quad \bigcap_k T'_k = \emptyset.$$

Let V be the maximum capability required for any AM $j \in N$ and for all CE $k \in M$.

The following definitions represent typical quantities of interest for any servicing operation:

$\phi_{1k}(X_i)$ Travel hours incurred by CE $k \in m_i$.

$\phi_{2k}(X_i|L_{v_1}^*)$ Overtime hours incurred by CE $k \in m_i$ (cumulative over v_1).

$\phi_{3j}(X_i)$ Response time incurred at AM $j \in n_i$, i.e., the sum of the repair request queuing delay at time i, the time required by the CE assigned to AM j to complete his current assignment and the CE's travel time to node j.

$\phi_{4j}(X_i|L_{v_2}^*)$ Observed probability over v_2 that a call originated by AM $j \in T'_k$ is *not* taken by CE k (prime CE).

$\phi_{5k}(X_i)$ Distance from CE k to T_u, where $k \in N^u$.

$\phi_{6k}(X_i|L_{v_3}^*)$ Cumulative time over v_3 that CE k is absent from T_u, where $k \in N^u$.

$\phi_{7k}(X_i)$ Difference between V and the capability measure for CE k (in terms of repair type and training).

In addition we define a corresponding group of prespecified parameters (with the argument suppressed):

a_{2k} Maximum overtime allowed for CE k (over v_1), $k \in m_i$.

a_{3j} Maximum required response time for AM j, $j \in n_i$.

a_{4j} Specified probability that prime CE will *not* visit AM j, $j \in n_i$.

a_{5k} Maximum distance from territory T_u for CE k, $k \in N^u$, $k \in m_i$.

a_{6k} Maximum time away from territory T_u for CE k (over v_3), $k \in N^u$, $k \in m_i$.

a_{7k} Minimum capability required for AM j, $j \in n_i$.

This notation defines the typical operating factors of interest such as the response time between node i and node j, i.e., the queuing time for AM j plus the time

required by the CE to complete service at i and travel to j, the effective cost of travel, cumulative overtime hours, time spent outside assigned territory, frequency of calls on subterritory by primary serviceman, extraterritorial travel and capability of the CE.

To simplify the notation we let X_i^* be the argument of the preceding functions. Thus, for constraints $2k$, $4j$ and $6k$, X_i^* is equivalent to $X_i|L_{v_1}^*$, $X_i|L_{v_2}^*$ and $X_i|L_{v_3}^*$, respectively, while for all other constraints $X_i^* = X_i$. Let $\psi_{p\,i}\phi_{p\,i}(X_i^*)$ be the cost associated with $\phi_{p\,i}(X_i^*)$. Further, let $H(t)$ and $G(t)$ be, respectively, the matrices of probability distribution functions for repair requests entering the system and completed on-site service activities (i.e., servicemen leaving the machine locations); the element in row k and column j corresponds to AM j serviced by CE k.

Mathematical program

The following mathematical program can now be formulated:

$$\min_{X_i} \{\mathrm{E}[\sum_{k\in m_t} \psi_{1k}\phi_{1k}(X_i^*) + \sum_{k\in m_t} \psi_{2k}\phi_{2k}(X_i^*)]\} \tag{1}$$

subject to

$$\phi_q(X_i^*) \leq a_q,\ q \in Q, \text{ for all } j \in n_t,\ k \in m_t, \tag{2}$$

where E denotes expectation with respect to the appropriate elements of $H(t)$ and $G(t)$ and $Q = \{(2, k), (3, j), (4, j), (5, k), (6, k), (7, k)\}$. Because X_i^* is an a priori decision made with respect to observation of the random variables associated with $H(t)$ and $G(t)$, infeasibilities in the constraints (2) can occur a posteriori. We avoid this problem by specifying a new objective function that includes the original costs plus the penalties for violating the constraints (2). Define

$$T^* = \{X_i^* : \mathrm{P}[\phi_q(X_i^*) \leq a_q] \geq \gamma_q\},\quad q \in Q,$$
$$\text{for all}\quad j, k;\quad 0 \leq \gamma_q \leq 1; \tag{3}$$

here P denotes probability. The appropriate penalty cost for violation of (2) can be established by first specifying Eq. (4); let

$$y_q = \max[0, \phi_q(X_i^*) - a_q]. \tag{4}$$

We reformulate the stochastic programming problem in a combined format,[2] obtaining

$$\min_{X_i^* \in T^*} \{\mathrm{E}[\sum_{k\in m_t} \psi_{1k}\phi_{1k}(X_i^*) + \sum_{k\in m_t} \psi_{2k}\phi_{2k}(X_i^*) + \sum_{\substack{q\in Q \\ \text{for all } j,k}} \psi_q(y_q, \bar{\alpha}_q)]\}, \tag{5}$$

where $\bar{\alpha}_q$ is the parameter vector to be determined. The stochastic program now includes the losses for violation of constraints in the objective function. The decision vector X_i is chosen from T^* such that the constraints are satisfied as chance constraints with a probability of feasibility given by the vector γ (with qth element γ_q). An adaptive control method is devised to accomplish this. The PACE dispatch system then adaptively determines a vector of parameters such that $X_i^* \in T^*$.

As a first-order approximation to the system, we determine the allocation of CE's to repairs that minimizes the expected immediate cost for a given parameter vector $\bar{\alpha}_q$. For the first-stage approximation, then, Eq. (5) can be written for fixed $\bar{\alpha}_q$ at $t = i$ as

$$\min_{X_i'} [\sum_{k\in m_t} \psi_{1k}\phi_{1k}(X_i') + \sum_{k\in m_t} \psi_{2k}\phi_{2k}(X_i') + \sum_{\substack{q\in Q \\ \text{for all } j,k}} \psi_q(y_q', \bar{\alpha}_q)] \tag{6}$$

subject to

$$X_i' \cdot E_1 = E_2 \quad \text{and}$$
$$E_2^T \cdot X_i' \leq E_1,$$

where $j \in n_i^*$, $k \in m_i^*$, m_i^* is the set of CE's available at $t = i$ and n_i^* is the set of AM's requiring service at $t = i$; also, X_i' is an $m_i^* \times n_i^*$ matrix in which $(X_i')_{kj} = 1$ if CE k is assigned to AM j but is 0 otherwise, y_q' is the value of y_q for $X_i^* = X_i'$ and E_1 and E_2 are n_i^*- and m_i^*-dimensional vectors of ones, respectively. The Balinski-Gomory algorithm[1] has been programmed and is currently being used to solve the modified assignment problem (6). Let ${}_0X_h'$ be the solution of (6) at time $t = h$. After observation of the system at $t = h, h+1, \cdots, h+k$, we determine the observed probability $\bar{\gamma}_q$ for satisfying constraint q corresponding to ${}_0X_h'$, ${}_0X_{h+1}'$, $\cdots$, ${}_0X_{h+k}'$. For the qth constraint the observed system error is defined to be

$$\epsilon_q = \max(0, \gamma_q - \bar{\gamma}_q).$$

Further, define U_q as the parameter-index set for the qth constraint penalty function and let $\bar{\alpha}_q$ be the parameter vector whose jth element is α_{qj}, $j \in U_q$. Parameter adjustment is accomplished by solving the auxilliary problem (7) to minimize the expected error,

$$\min_{\bar{\alpha}_q} \mathrm{E} \sum_q \epsilon_q(\bar{\alpha}_q), \qquad q \in Q. \tag{7}$$

- *Constraint combination*

To reduce the control parameter space, we let the performance index for each CE and each account be identical, i.e., in (3) set

$$\gamma_q = \gamma_q^* \quad \text{and} \tag{8}$$
$$a_q = a_q^*,\ q \in Q \text{ for all } j \in n_t^*,\ k \in m_t^*. \tag{9}$$

Further, we let the constraints in (3) be satisfied in an aggregate sense; these can now be written as

$$F_q(a_q^*) \geq \gamma_q^*, \qquad q \in Q^* \quad \text{and} \quad 0 \leq \gamma_q^* \leq 1, \tag{10}$$

where $F_q(a_q^*) \equiv \mathrm{P}\,[\phi_q(X_i|L_v^*) \leq a_q^*]$ and Q^* is the index set $\{2, 3, 4, 5, 6, 7\}$. The auxillary parameter control problem (7) then becomes

$$\min_{\bar{\alpha}_q} \mathrm{E} \sum_q \epsilon_q(\bar{\alpha}_q), \qquad q \in Q^*. \tag{11}$$

• *Adaptive parameter adjustment*

We solve control problem (11) by designing an adaptive system based on the essential concepts of feedback control and forecasting. The optimal values of α_{qj} at any time t are obtained by estimating statistically the expected total error $\mathrm{E} \sum_{q \in Q^*} [\epsilon_q(\alpha_{qj})]$ as a function of α_{qj} during a predetermined interval of time $[t, t + \Delta t]$. The interval Δt is selected so that the probability-distribution-function matrices $H(t)$ and $G(t)$ can be assumed to be constant during this interval. The value of Δt also depends on the variability of the load and service conditions of the system, but need not be small.

At time t let $\alpha_{qj}(t)$ denote the value of the jth parameter α_{qj}, where $j \in U_q$, U_q being the parameter-index set for the qth constraint penalty function as defined previously. After observing the system with this value of the parameter at various points in the interval $[t, t + \Delta t]$, during which n assignments are made, we note the performance of the system at time $t + \Delta t$. Violations of some of the constraints (10) can, of course, have occurred during this interval. Define $I_0(t)$ to be the set of indices of these constraints, i.e., $I_0(t) = \{q : q \subset Q^*, F_q(a_q^*) < \gamma_q^*\}$. We now adjust sequentially (following a prespecified order) the parameter(s) α_{qj} to compensate the constraint violations as follows:

$$\alpha_{qj}(t + \Delta t) = \alpha_{qj}(t) + \Delta\alpha_{qj}(t + \Delta t),\ q \in I_0, \text{ all } j \in U_q, \tag{12}$$

where $\Delta\alpha_{qj}(t + \Delta t)$ is obtained from the response surface of $F_q(a_q^*)$ with respect to parameter(s) α_{qj}. A new target $\tau_q(t + \Delta t)$ based on the performance of n assignments during the interval $[t, t + \Delta t]$ is calculated. The interpretation of $\tau_q(t + \Delta t)$ and its calculation are discussed later. The *initial* values of the parameters are arbitrary and can be based on judgment guided by such factors as (a) the relative values of each penalty between penalty classes and (b) the specific losses associated with deviations from the desired performance. Assuming temporarily that $\tau_q(t + \Delta t)$ is available, we examine the relation between $F_q(a_q^*)$ and the control parameters. For convenience, we refer to this relation as the *response surface* of $F_q(a_q^*)$ and we show how the adjustment $\Delta\alpha_{qj}(t + \Delta t)$ is obtained using the target values and the response surface.

• *Response surface*

The response surface of $F_q(a_q^*)$ can be written as

$$F_q(a_q^*) = R_q(\bar{\alpha}_q, t), \qquad q \in Q^*, \tag{13}$$

where $\bar{\alpha}_q$ is a vector whose elements are α_{qj}. The functional form of the response surface of $F_q(a_q^*)$ is not obtainable analytically and the unknown $R_q(\bar{\alpha}_q, t)$ is a time-dependent function. However, $F_q(a_q^*)$ can be evaluated for given parameter values $\bar{\alpha}_0$ by observing the performance of the system. To construct a response surface, simultaneous observation of $F_q(a_q^*)$ for more than one value of $\bar{\alpha}_q$ is necessary. This is difficult in an operating system. The dynamic nature of R_q is due to both inter- and intra-day changing load and service conditions. The alternative is to develop estimates of $F_q(a_q^*)$. An obvious, though infeasible, way of doing this is to simulate the operation of the system during real-time operation. In the present PACE dispatch system this would require excessive computing time. Therefore we obtain these estimates indirectly as follows:

Consider the assignment matrix at time $t + \Delta t$. It contains the status of all CE's such that $m \in m_t^*$, i.e., those working or free during $[t, t + \Delta t]$. Let $z_{qmn}(t + \Delta t)$ be the estimated value of the operating factors, also referred to as the system output, corresponding to the desired value a_q^* if CE $m \in m_t^*$ were assigned to account $n \in n_t^*$. Denote the optimal solution using parameter value(s) $\alpha_{qj}(t)$ by $X(t + \Delta t)$ and define $z_{qmn}^*(t + \Delta t) = z_{qmn}(t + \Delta t)$ for m, n such that $X_{mn}(t + \Delta t) = 1$. Thus $z_{qmn}^*(t + \Delta t)$ represents the value of the output if CE $m \in m_t^*$ in the optimal solution $X(t + \Delta t)$ were actually dispatched to the assigned account $n \in n_t^*$. Let the fitted distribution of the estimated output values over all m and n in $X(t + \Delta t)$ be $F_q^*[z_q^*(t + \Delta t)|\alpha_{qj}(t + \Delta t)]$; then, for large numbers of CE's and accounts, F_q^* is found to be constant during interval Δt (not necessarily small) in which the load and service conditions of the system do not change significantly. Because not all of the CE's considered for assignment are free at time t, they are not all dispatched; dispatching is performed only for free CE's. Assuming that each CE is equally likely to be free during $[t, t + \Delta t]$, we can consider the observed value of

$$\bar{\gamma}_q[(t + \Delta t)|\alpha_{qj}(t + \Delta t)]$$

for the actually dispatched CE's as selected randomly from the distribution $F_q^*[z_q^*(t + \Delta t)|\alpha_{qj}(t + \Delta t)]$ and we can use $F_q^*[z_q(t + \Delta t)|\alpha_{qj}(t + \Delta t)]$ for estimating this distribution in Eq. (13). Alternate optimal assignments corresponding to s different values $\alpha_{qj}^k(t + \Delta t)$, $k \leq s$, of the parameter(s) α_{qj} suggested by the next target value τ_q are simultaneously determined. Estimates of $F_q[a_q^*|\alpha_{qj}^k(t + \Delta t)]$ are obtained by calculating $F_q^*[z_q^*|\alpha_{qj}^k(t + \Delta t)]$, $k \leq s$, from the observed values of z_q^*.

The number of points s required to construct a good response surface and the selection of response surface equations for each constraint depend on the nature of

the response surface, the number of parameters and the degree of interaction among the parameters. The response surface can be found best experimentally. Later in this paper we establish that, under certain fairly nonrestrictive conditions which are satisfied by the system, the error ϵ_q is a monotonically decreasing function of each parameter corresponding to a constraint. Because

$$\epsilon_q = \max [\gamma_q^* - F_q^*(a_q^*), 0]$$

and γ_q^* is a specified constant, the response surface of F_q^* is a monotonically increasing function of each parameter. In addition, evidence indicates that, in practice, sufficient control can be exercised by adjusting only the most sensitive parameter, in which case appropriate linear or nonlinear functions can be fitted easily. Techniques for the estimation of nonlinear parameters[3,4] can be used in cases of two or more interacting parameters.

If the probability of being free during $[t, t + \Delta t]$ is not the same for each CE, F_q^* turns out to be a biased estimator of $F_q(a_q^*)$. In such a case experimental observation of this bias improves the accuracy of the fitted response surface. Let the response surface for constraint q near parameter value $\alpha_{qj}(t)$ be represented by

$$F_q[a_q^* | \alpha_{qj}(t + \Delta t)] = h[\alpha_{qj}(t), \Delta\alpha_{qj}(t + \Delta t)].$$

If the new target probability at time $t + \Delta t$ is $\tau_{qj}(t + \Delta t)$, $\Delta\alpha_{qj}(t + \Delta t)$ is obtained by solving

$$\tau_{qj}(t + \Delta t) = h[\alpha_{qj}(t), \Delta\alpha_{qj}(t + \Delta t)].$$

After the necessary adjustment is made in the parameters corresponding to constraint q, the response surface for the next constraint in sequence, $q^* \in I_0$, is obtained similarly during the next assignment. This response considers the effect of adjustment in the previous parameter(s) because the effect of adjustment on the system output is immediate. In addition, the sequential adjustment procedure distributes the computing time over several assignments. This is an important consideration in real-time operating systems.

In adjusting $\Delta\alpha_{qj}(t)$ sequentially, the effects of parameter changes are considered one at a time, with each step corresponding to one constraint. The interaction among the constraints is ignored. This is not serious in practice for three reasons: (a) Knowledge of the behavior of the system allows recognition of those constraints that are highly interactive and indicates the direction of their interaction, i.e., in (13) the components $\alpha_{q'j'}$, $q' \in Q^*$ and $j' \in U_q$, with the most significant values of signed partial derivatives, $\partial F_q(a_q^*)/\partial\alpha_{q'j'}$, $q' \neq q$, are indicated; (b) the direction of interaction among the constraints is time invariant, i.e., the algebraic sign of the above partial derivative does not change with time; and (c) the effect on the remaining parameters of the adjustment of one parameter is implicitly accounted for when the next constraint in sequence is considered. A "least" interactive sequence can be determined with knowledge of (a) and (b).

• *Target values*

For the previous N_q assignments and the qth constraint we observe $F_q(a_q^*)$. Satisfactory performance is obtained only when $F_q(a_q^*) \geq \gamma_q^*$. Number these assignments in the order of their execution, i.e., 1, 2, $\cdots$, N_q, and let $z_1, z_2, \cdots, z_{N_q}$ be the corresponding observed values of the output. Find the shortest sequence $z_1, z_2, \cdots, z_n$, $n \leq N_q$, for which $F_q(a_q^*) \geq \gamma_q^*$. Drop these n observations and renumber the remaining ones as $z_1, z_2, \cdots, z_{N_q - n}$. Repeat this elimination procedure until n_q (≥ 0) observations are left for which $F_q(a_q^*) < \gamma_q^*$. If $n_q = 0$ the target value is γ_q, but if $n_q > 0$ proceed as follows: If the adjusted value of the parameter is used during the interval $[t, t + \Delta t]$ and the estimated number of assignments during this period is m_q, to obtain $F_q(a_q^*) \geq \gamma_q^*$ for $n_q + m_q$ observations, γ_q^* should be obtained such that $F_q(a_q^*) \geq \gamma_q' > \gamma_q$ for the next m_q observations. The 90%-confidence estimate of γ_q', for example, can be obtained by solving

$$\sum_{i=v}^{m_q} \binom{m_q}{i} (\gamma_q')^i (1 - \gamma_q')^{m_q - i} \leq 0.10, \tag{14}$$

where v is the maximum number of violations that can be permitted in the next m_q observations such that $F_q(a_q^*) \geq \gamma_q^*$. If at the end of the next m_q observations $F_q(a_q^*) \geq \gamma_q^*$, then $n_q = 0$ by the elimination procedure and the next target value is γ_q^*; if not, i.e., if $n_q > 0$, the new target value can be estimated using (14).

The advantage of using this procedure rather than using all N_q observations is that the target value is automatically kept constant and no manipulation on the previous N_q observations is required when the constraint q is satisfied. The procedure is also designed to maintain the probability of violations in N_q observations within the feasible region.

• *Penalty functions*

A field engineering study was made to ascertain the utility functions associated with losses incurred for the various error conditions in the proposed constraints.[5,6] The results of this study provided confidence in using the class of exponentials as a class of loss functions representing best current field engineering practice. Selection of a class of loss functions consistent with current judgment provides a conservative approach to selection of these functions. However, we do not specify the particular parameters associated with the exponential penalties, but obtain these parameters through an adaptive procedure for optimal performance as defined by a performance index.

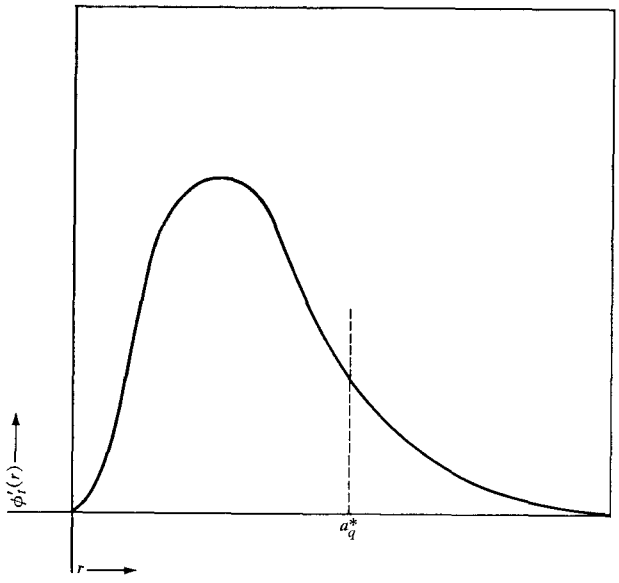

Figure 3 Distribution of estimates of $z_{qmi_0}(t)$.

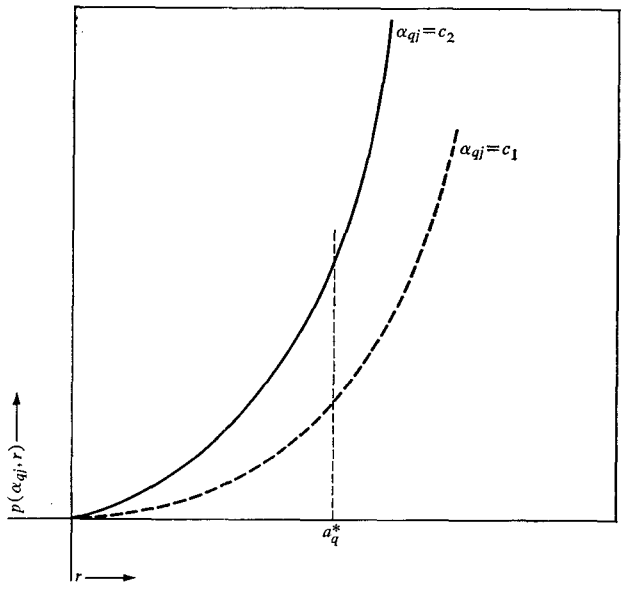

Figure 4 One-sided penalty function for α_{qj}.

The most general form of the exponential penalty function corresponding to the qth constraint and the decision X_t can be written as

$$p_{qkj}^t = \alpha_{q1}[e^{\alpha_{q2}(z_{qkj}^+ - \alpha_{q3})} - 1] + \alpha_{q4}[e^{\alpha_{q5}(z_{qkj}^- - \alpha_{q3})} - 1] + \alpha_{q6}z_{qkj}^+ + \alpha_{q7}z_{qkj}^-,$$

where

$$z_{qkj}^{\pm} = \max\{0, \pm[z_{qkj}(t) - a_{qj}]\}$$

and α_{qi}, $i \in U_q$, are the control parameters corresponding to the qth constraint. The exponential function p_{qkj}^t is the penalty corresponding to constraint q and decision X_t associated with assigning CE k to account j. All the components, i.e., linear and exponential, need not be present in the penalties; the components used were determined by the field engineering study.[5] For example, consider the constraint corresponding to response time in definition (3). The satisfaction of this constraint is specified in terms of the desired response time a_{3j} and $F_3(a_{3j}) \geq \gamma_3$. The penalty function used is

$$p_{3kj}^t = \begin{cases} \alpha_{31}[e^{\alpha_{32}[z_{qkj}(t) - \alpha_{33}]} - 1], & z_{qkj}(t) \geq \alpha_{33}, \text{ or} \\ \alpha_{34}[e^{\alpha_{35}[\alpha_{33} - z_{qkj}(t)]} - 1], & z_{qkj}(t) < \alpha_{33}. \end{cases}$$

The mean of the output distribution increases (decreases) as the value of α_{33} increases (decreases), whereas the criterion $F_3(\alpha_{3j}) \geq \gamma_3$ can be achieved by adjusting appropriately the values of α_{32}, α_{33} and α_{35}. Note that a change in the values of α_{31} and α_{34} changes this penalty cost relative to other penalties.

• *Monotonic property of the error*

For the PACE dispatch system, in which exponential penalty functions are used and the appropriate algebraic sign is assigned to each parameter in the penalty function, the response surface of the error ϵ_q, $q \in Q^*$, for each individual constraint q is a monotonically decreasing function of each parameter α_{qi}, $j \in U_q$. We next establish this monotonic property for a more general class of functions and state the conditions under which it holds.

Currently, a set $\{z_{qmn}(t)\}$ of estimates of the expected values of the output variables is used to determine the optimal assignment at time t. Consider one variable $z_{qmn}(t)$, $q \in Q^*$, $m \in m_t$ and $n \in n_t$, for a specific account $l_0 \in n_t$. For convenience, let $r_1(t)$, $r_2(t)$, $\cdots$, $r_p(t)$ represent the set of estimated expected values of $z_{qml_0}(t)$ arranged in an increasing order, i.e., $r_1(t) \leq r_2(t) \leq \cdots \leq r_p(t)$, for p CE's qualified and available to answer a call from this account. Consider the effect of just one parameter α_{qi}, $j \in U_q$. Figure 3 shows the distribution $\phi_t'(r)$ of the set of these p estimates and Fig. 4 shows the penalty function, here denoted by $p(\alpha_{qi}, r)$, corresponding to this constraint.

Let $F_r[x|c_1, \phi_t'(r)] = \mathrm{P}[r \leq x|c_1, \phi_t'(r)]$, the conditional probability that the resulting value r will be less than or equal to x if the parameter $\alpha_{qj} = c_1$ and if the density function of the estimated r is $\phi_t'(r)$. Also, let $[p_{il_0}|c_1, \phi_t'(r)]$, $i = 1, 2, \cdots, p$, be the probability of CE i's being assigned to the account l_0 under consideration and let $[r^*|c_1, \phi_t'(r)]$ be the resulting observed value at time t. This latter value depends on the distribution $F_r'[x|c_1, \phi_t'(r)]$ through $[p_{il_0}|c_1, \phi_t'(r)]$ and $\phi_t'(r)$.

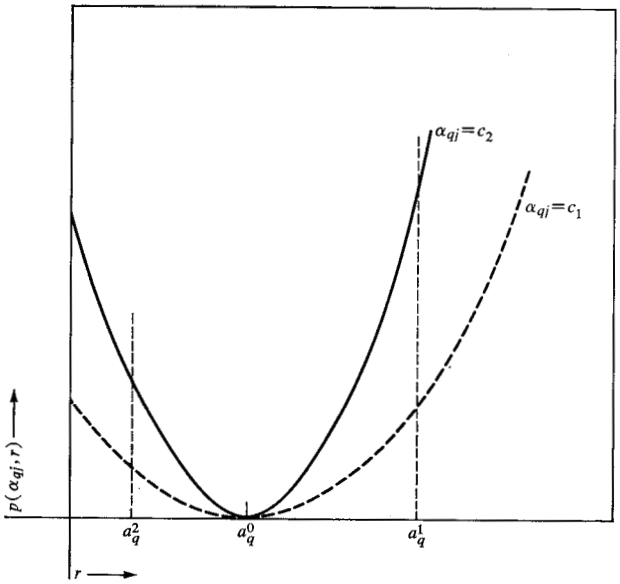

Figure 5 Two-sided penalty function for α_{qj}.

Now, if an incremental cost $\Delta c_{i l_0} = \Delta c$, $i = 1, 2, \cdots, p$, is added to column l_0 of the cost matrix (of the assignment problem), for a cost matrix with random c_{ij} the quantity $[p_{i l_0} | c_1, \phi'_{n+1}(r)]$ is unchanged. However, if

$$\Delta c_{1 l_0} < \Delta c_{2 l_0} < \cdots < \Delta c_{p l_0} \tag{15}$$

as is the case when α_{qj} is replaced by $\alpha_{qj} + \Delta\alpha_{qj}$ in the penalty functions of exponential form, the probability $[p_{i l_0} | c_2, \phi'_i(r)]$, where $c_2 = \alpha_{qj} + \Delta\alpha_{qj}$, of assigning the CE having the appropriate set of characteristics denoted by $r_{i l_0}(t)$ will increase relative to the probability $[p_{i+1, l_0} | c_2, \phi'_i(r)]$ for the CE with $r_{i+1, l_0}(t) > r_{i l_0}(t)$. In other words,

$$F_r[x \mid c_2, \phi'_i(r)] \geq F_r[x \mid c_1, \phi'_i(r)]; \tag{16a}$$

consequently

$$F_r[a_q^* \mid c_2, \phi'_i(r)] \geq F_r[a_q^* \mid c_1, \phi'_i(r)], \quad c_2 > c_1, \tag{16b}$$

and

$$[r^* \mid c_2, \phi'_i(r)] \leq [r^* \mid c_1, \phi'_i(r)], \quad c_2 > c_1, \tag{16c}$$

where a_q^* is the upper limit on the desired value of the output variable.

This argument can be extended to time $t + \Delta t$ by assuming that $\phi'_{t+\Delta t}(r)$ remains unchanged despite any change in assignments. If the number of CE's in the system qualified to service acount l_0 is large, we can assume that $\phi'_i(r) = \phi'(r)$ during period $[t, t + \Delta t]$, i.e., that this distribution does not change significantly as a result of changes in assignments due to an increase or a decrease in the value of α_{qj}. Experimental results for the PACE dispatch system support this assumption. Thus we can write

$$\epsilon_q(c_2) \leq \epsilon_q(c_1), \qquad c_2 > c_1,$$

because

$$\epsilon_q(\alpha_{qj}) = \max \{\gamma_q - F_q[a_q^* | \alpha_{qj}, \phi'_i(r)], 0\}$$

and $(r^* | c_2) \leq (r^* | c_1)$, $c_2 > c_1$, during the interval $[t, t + \Delta t]$ when $\phi'(r)$ does not change.

For the penalty functions shown in Fig. 5, where the error is defined as

$$\epsilon_q(\alpha_{qj}) = \max \{\gamma_q - \langle F_q[a_q^1 \mid \alpha_{qj}, \phi'_i(r)] - F_q[a_q^2 \mid \gamma_{qj}, \phi'_i(r)]\rangle, 0\},$$

a_q^1 and a_q^2 are upper and lower limits, respectively, on the desired output value and γ_q is the specified minimum for the probability that the observed values fall between these limits. A similar argument can be used to derive

$$\epsilon_q(c_2) \leq \epsilon_q(c_1),$$

$$(r^* \mid c_2) \leq (r^* \mid c_1), \quad c_2 > c_1, \quad r^* > a_q^0,$$

and

$$\epsilon_q(c_2) \leq \epsilon_q(c_1),$$

$$(r^* \mid c_2) \geq (r^* \mid c_1), \quad c_2 > c_1, \quad r^* \leq a_q^0,$$

where a_q^0 is the value of the output variable such that $p(\alpha_{qj}, a_q^0)$ is the minimum penalty.

In general, for (15) to be satisfied the condition required on the penalty function is given by Eqs. (17). Let a_q^0 be defined as above; then

$$\min [p(\alpha_{qj}, r)] = p(\alpha_{qj}, a_q^0). \tag{17a}$$

This implies that

$$p(\alpha_{qj}, r^1) \geq p(\alpha_{qj}, r^2) \tag{17b}$$

if $r^1 > r^2 \geq a_q^0$ or if $r^1 < r^2 < a_q^0$ and

$$p(\alpha_{qj}^1, r) \geq p(\alpha_{qj}^2, r)$$

if $\alpha_{qj}^1 > \alpha_{qj}^2$. (17c)

The argument can be extended to each parameter α_{qj}, $j \in U_q$. Denoting the observed output value for constraint q at time t by $z_q(t)$, we can, in general, write

$$[z_q(t) \mid \alpha_{qj}(t)] \leq [z_q(t) \mid \alpha_{qj}^1(t)] \tag{18a}$$

for $\alpha_{qj}(t) > \alpha_{qj}^1, z_q(t) \geq a_q^0$, and

$$[z_q(t) \mid \alpha_{qj}(t)] \geq [z_q(r) \mid \alpha_{qt}^1(t)] \tag{18b}$$

for $\alpha_{qj}(t) > \alpha_{qj}^1, z_q(t) < a_q^0$, which gives

$$\epsilon_q[\alpha_{qj}(t)] \leq \epsilon_q[\alpha_{qj}^1(t)] \tag{19}$$

for $\alpha_{qj}(t) > \alpha_{qj}^1(t)$ in the interval $[t, t + \Delta t]$. The validity of condition (18) rests on the assumption that $\phi'_i(z_q)$ remains unchanged during $[t, t + \Delta t]$ and that relations (17) hold.

Summary

With the increasing trend toward automation and more sophisticated machinery, the service function assumes an ever increasing importance. PACE dispatch has been designed as a real-time system to assign customer engineers (servicemen) to requests for service, preventive maintenance and engineering- and sales-change activities. The system is intended to allow management to make decisions on factors such as overtime and response time and to be assured that commitments on those factors will be met. The adaptive features incorporated in PACE dispatch were designed to provide a system that would operate satisfactorily with IBM Field Engineering Division branch offices which vary both geographically and in service-force structure. In addition, treating the service problem in the framework of a stochastic programming formulation is consistent with field engineering practices and provides a system that can be incorporated into the Field Engineering Division operation without a major restructuring of current methods. Tests based on data collected during a one-month period in both the Brooklyn, New York and Washington, D. C. branch offices indicate that the PACE dispatch system could handle 11 to 18 percent more workload than the current system while maintaining equivalent or better levels of service.[5,6]

Acknowledgments

We acknowledge the efforts of B. K. Knoppers of the University of Maryland for the programming of PACE Dispatch I and the programming assistance of O. Mond with PACE Dispatch II. In addition, the cooperation and enthusiasm of the Field Engineering Division in providing programming assistance under C. F. Weiss, Jr. to field test this system are appreciated.

References

1. M. L. Balinski and R. Gomory, "A Primal Method for the Assignment and Transportation Problems," *Management Sci.* **10,** 578 (1964).
2. W. H. Evers, "A New Model for Stochastic Linear Programming," *Management Sci.* **13,** 680 (1967).
3. Y. Bard, "Nonlinear Parameter Estimation and Programming," *SHARE Program 360D-13.6.003,* 1967.
4. D. W. Marquardt, "An Algorithm for Least-Square Estimation of Nonlinear Parameters," *SIAM J. Appl. Math.* (formerly *J. Soc. Indust. Appl. Math.*) **11,** 431 (1963).
5. W. H. Evers, "PACE Dispatch Final Report I," *Operations Research Report CES-FSD 10667001,* IBM Federal Systems Division, Gaithersburg, Maryland, October 1967.
6. W. H. Evers and B. K. Knoppers, "PACE Dispatch Final Report II," *Operations Research Report CES-FSD 121567002,* IBM Federal Systems Division, Gaithersburg, Maryland, December 1967.

Received January 29, 1969