DEDICATION VERSUS FLEXIBILITY IN FIELD SERVICE OPERATIONS

Fikri Karaesmen[†]
Frank Van der Duyn Schouten [††]
and
Luk N. Van Wassenhove[†††]

[†] Laboratoire d'Informatique de Paris 6 (LIP6-CNRS)
Université Pierre et Marie Curie
4, place Jussieu
75252 Paris Cedex 05, France
Fikri.Karaesmen@lip6.fr

[††] Center for Economic Research
Tilburg University
5000 LE, Tilburg, The Netherlands

[†††] INSEAD
Boulevard de Constance
77300, Fontainebleau Cedex, France

May 1998

Acknowledgements: Part of this research was realized when F. Karaesmen was visiting the Center for Economic Research of Tilburg University. The author thanks CentER for making this visit possible. We thank Onno Boxma, Harrie de Haas and Ger Koole for helpful discussions and pointing out relevant references.

Abstract

Field service is gaining importance as after sales service is starting to be recognized as a major source of revenue. This motivates planning problems for companies that employ mobile technicians who provide service on clients' sites. These planning problems share the common characteristic that service levels corresponding to technician response times are explicitly expressed in contracts. Moreover, lately, there is strong pressure from clients to have a single dedicated technician who takes full responsibility of the field service. In this paper, we provide models that enable the analysis of various trade-offs between service levels and operational costs under the dedicated service structure. We also investigate the tradeoffs between strict dedicated service and more flexible structures to understand the settings for which strict dedication is appropriate.

1 Introduction

The importance of field service is rapidly growing as after sales service quality has become a significant portion of product offerings. In their benchmarking study of after-sales service logistics systems in the computer industry, Cohen, Zheng and Agrawal (1997) state: "Customer expectations for product reliability have increased. As a result, the provision of superior after-sales service, at a competitive price, has become an important qualifier for competitive survival". Increasingly, companies take a life cycle approach to their products realizing that lower sales margins can often be more than compensated by lucrative long term service contracts. From our personal contacts, we know that companies like Otis, Xerox, ABB, SKF, GE and GEC-Alsthom make a significant and rapidly increasing part of their total revenues in after sales activities. In Cohen et al.'s sample of the computer industry after sales revenues were, on average, equal to 30 % of product sales (Cohen, Zheng and Agrawal, 1997).

In this paper, we consider operational level planning problems for companies that provide repair and maintenance to clients' equipment via mobile technicians. Typically the *clients* that are considered in our framework are holders of multiple equipment at a single site to be serviced by the firm. Although

the application that has motivated this research is the case of automated vending machines, problems of similar nature are abundant in after sales service by manufacturers of high-technology equipment (photocopy machines, printers, computers, etc.). As an example, Cohen, Zheng and Agrawal (1997) report that the average computer firm in their sample had about 100000 installed machines, 300 service engineers and 65000 service calls per year.

Firms compete in field services through quality of the service they provide. Unfortunately, service quality depends heavily on the perceptions of the client and objective service level measures are difficult to set. In the field service support setting, a critical measure that shapes the perception of the client is the response time of the firm in the case of unforeseen breakdowns of equipment. Hence, one can argue that clients have a preference for firms that provide shorter average response times in the long run. In addition, in the shorter term, to establish their competitive advantage, firms provide contractual agreements with their clients that specify guarantees based on response time measures. However, these guarantees are in terms of response time limits rather than average response times. For example, the contract may specify that 90 % of the service requests are met within 8 hours of the request. For example, many companies in the computer industry set their service target by specifying the percentage of customer demands that are met within 24 hours. For critical applications, service providers are required to guarantee service within two hours of product failure. Some companies offer service guarantees between the two-hour and the 24-hour standard (i.e., 8 or 12 hours) (Cohen, Zheng and Agrawal, 1997).

A second important aspect that defines service quality from the clients' side is the interaction with the firm. A common request from the client side is to have a technician assigned to the client who is entirely responsible for the repair and maintenance of the equipment at that site. In other words, clients have a strong preference to deal with a unique account representative rather than having service provided by a different technician each time. This is a burdensome request for the service firm; it is well known that to minimize costs, each client has to be served by a pool of technicians rather than a single

account representative.

In this paper, we provide planning models that combine the two important issues mentioned above: response time guarantees and unique account representatives. We first provide a framework to measure the performance of a single account representative assigned to a number of client sites. We then determine the number of account representatives required to handle the servicing of a number of sites with response time constraints under the account representative setting. Finally, we provide a model that measures the performance difference between a strict account representative setting and a more flexible setting in which clients are served by their representatives most (but not all) of the time.

We are aware of only a few papers that directly study field service design issues. Smith (1979) presents a queueing model to estimate the territory size that can be covered by a single service representative when service requests are distributed uniformly within the territory. He shows that the response time performance measures in the model are equivalent to those in a corresponding M/G/1 queue. Hill et al. (1992a) consider the case of Smith's model with multiple servers per territory and give an approximating M/G/c type queueing model. A key observation is that response times deteriorate as the number of busy servers increase (due to a larger distance between the available server and the place from which the service request originates), hence the appropriate approximating model is an M/G/c queue with service rates dependent on the number of busy servers. Hill (1992b) studies dispatching rules for multiple technicians responding to service requests in a territory. Through extensive simulation experiments he shows that a first come first serve dispatching rule performs poorly and proposes dispatching rules that combine travel time considerations with delay limit considerations.

From a more practical perspective, Hambleton (1982) describes the issues that have to be considered for a field service firm that maintains vending machines in England. A hierarchy of problems is described. At the highest level, the size of each separate service region and its approximate capacity is determined. At the medium level, within each region, a *patch* of customers is

allocated to each technician. Finally, at the lower level, detailed scheduling decisions take place.

Apart from the models that directly deal with field services, another related class of problems are those based on dynamic vehicle routing. Bertsimas and Van Ryzin (1993) study a dynamic vehicle routing problem with multiple vehicles. In particular for demands uniformly distributed in a region, they analyze routing policies to minimize expected waiting time costs. Dynamic vehicle routing models are appropriate for services where transportation times are significant in comparison to actual on-site service times. Our models fall outside of this category as we consider cases with significant on-site service times. In fact, we do not attempt to analyze the effects of sequencing of service (the routing of technicians) in our model. One can argue that there is not much room for sequencing with tight response time limits and a dense geographical region, so, the effects of sequencing are not as critical for the models that we consider.

The account representative-client site assignment problem has the flavor of other assignment problems for service system design. For example, the garbage truck-dump site assignment problem studied by Agnihothri, Narasimhan and Pirkul (1990) considers a service system design problem where clients have to be assigned to servers with expected waiting time considerations. Amiri (1997) extends this model to include servers with different capacity levels and provides a solution methodology. Melachrinoudis (1994) considers a version of the discrete location assignment problem with queueing effects.

Our model differs from those considered in the above papers in several ways. First of all, the starting point of all of the above models is that the service requests are uniformly distributed in a geographical region. In our case, we consider *client sites* that are fixed in location and that contain multiple machines. Secondly, we handle delay limits directly rather than considering average response times or variances of response times as in Smith (1979). This requires the use of more sophisticated tools recently developed for the analysis of multiclass queues. Thirdly, we focus on utilizing multiple technicians as account representatives assigned to sites rather than a pool as in Hill (1992a, 1992b).

The final difference between the above papers and ours is that they provide a formulation for a particular problem and develop a solution methodology for it, whereas we formulate a number of plausible problems to demonstrate how response time limit constraints can be incorporated in assignment formulations.

In section 2, we introduce the general framework by modeling a single account representative who serves a given set of clients. Section 3 uses the framework to develop tractable formulations for a variety of design problems that arise under the dedicated setting , i.e. when each client site is assigned a unique representative. In section 4 we develop a simple model to analyze the effects of *dedication with a degree of flexibility*. Using the model, we compute the difference in performance for different degrees of flexibility and under different service structures. Finally, we present our conclusions in section 5.

2 A Single Account Representative

In this section, we introduce basic models of a single account representative who serves a patch of clients. We first define the processes that describe client repair times. The travel times are more complicated as they are dependent on the service strategy of the representative, hence we discuss the modeling assumptions under different service strategies.

In our models a client is a site that contains multiple machines which are serviced by the firm. In general, a client is a single company which has a contract with the service provider. Naturally, one can lump a number of closely located small clients into a single client for modeling purposes. Under this assumption each client has a fixed site to which the associated account representative has to travel to perform the repair. As each client may have a different number and mix of equipment, the breakdown rates and repair times have to be client dependent.

To capture the above characteristics, we assume that an account representative services call requests from I sites in a region. A client i generates a call request (which corresponds to an equipment breakdown) at rate λ_i according to a Poisson process. Each on-site repair takes an amount of time that depends

on the characteristics of the mix of equipment at the site. We denote by R_i the random variable that models the on-site repair times. Next we discuss the travel strategies of the server and outline different models for different travel strategies.

First, we consider an account representative who has a fixed base location to which he has to return after each serviced call. We define by d_i the distance from site i to the base. Travel times of the server from the base to a site i are random variables T_i that depend on the distance d_i. For clarity, assume that trips from the base to a site i have the same distribution as return trips from site i to the base. Figure 1 displays this setup.

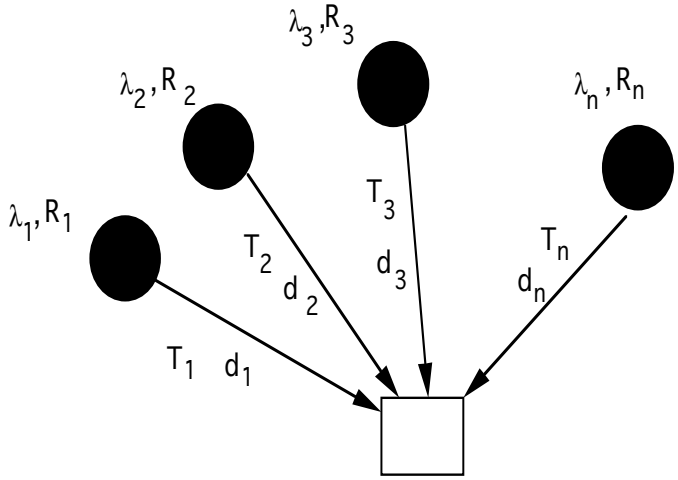

Figure 1: The setup for n sites assigned to a representative

Letting S_i denote the *total service time* requirement of site i, we have that S_i is composed of a travel time to site i, T_i, a repair time R_i at site i, and a return travel time, $\hat{T}_i$ (where T_i and $\hat{T}_i$ are independent and identically distributed random variables). Hence

$$S_i = T_i + R_i + \hat{T}_i \tag{1}$$

To complete the analogy to an M/G/1 queue let $\lambda = \sum_{i=1}^{I} \lambda_i$ denote the total call rate for service requests and let p_i define the long run proportion of

the total call rate that is initiated by site i, i.e:

$$p_i = \frac{\lambda_i}{\lambda} \tag{2}$$

Finally letting S be the mixture of the S_i with mixture weights p_i (i.e. the probability distribution function of S is a weighted sum of the probability distribution functions of S_i with respective weights p_i), the representative is the *server* in an M/G/1 queue with arrival rate λ and service time S. Hence, for this model, we can compute performance measures of interest by analyzing the corresponding M/G/1 queue.

In many field service environments, the representatives do not return to the base after each service completion but travel from one site to another sequentially. Consider a small geographical region that covers a large number of sites. A typical example of this case is a district in a large city. As the region is small and the number of sites is large, travel times are almost independent of the sequence of sites to be served within a trip of the technician and hence they can be modeled as independent and identically distributed random variables. Let T be this random variable that represents the common travel time.

When an account representative is assigned to this group of sites, we have the M/G/1 analogy, with total arrival rate $\lambda = \sum_{i=1}^{n} \lambda_i$, and with service time which is the mixture of $T + R_i$ with mixing weights p_i.

When the travel times are significantly different between sites, using a unique site-independent travel time random variable would be too crude an approximation.

Let d_{ij} denote the distance between site i and site j and T_{ij} denote the corresponding travel time random variable (with the understanding that $T_{ii} = 0$, for all $i = 1, 2, ..., I$). We assume that if there are no calls to be served at the end of a service completion, the representative stays at the last site and that calls are served according to a first-in-first-out discipline.

Let T_i denote the travel time to a site i conditioning on the site which generated the previous call, we obtain T_i as the mixture of T_{ji} with mixing weights p_j.

The total service time in this case consists of a travel time from the previous

site and a repair time and is given by:

$$S_i = T_i + R_i \tag{3}$$

Once again the representative is the server of an M/G/1 queue with arrival rate λ and service time S where the probability distribution function of S is a weighted sum (with weights p_i) of the probability distribution functions of S_i's.

The advantage of setting representative assignments as M/G/1 type models is the availability of tools for performance analysis. We discuss the tools for two performance measures that are of key interest to us: average delays and tail distributions for the delays.

Let S be the random variable denoting the service time, $S(x)$ its cumulative distribution function and W_q the delay (before service). The well-known Pollaczek-Khinchine formula gives the expected value of W_q:

$$\mathrm{E}[W_q] = \frac{\lambda \mathrm{E}[S^2]}{2(1-\rho)} \tag{4}$$

with $\rho{:=}\lambda \mathrm{E}[S] < 1$

The tail probabilities of W_q are much more difficult to obtain than its expected value. Using asymptotic expansions, Tijms (1994) provides:

$$\mathrm{P}\{W_q > w\} \approx \gamma e^{-\delta w} \tag{5}$$

where $\delta > 0$ is the unique root of:

$$\lambda \int_0^\infty e^{\delta x}(1 - S(x))dx = 1 \tag{6}$$

and

$$\gamma = (1-\rho)\left[\lambda\delta \int_0^\infty x e^{\delta x}(1 - S(x))dx\right]^{-1} \tag{7}$$

Recently, Abate, Choudhury and Whitt (1995) have given further theoretical and experimental support for the accuracy of the above approximation to compute tail probabilities in GI/G/1 queues.

The waiting time distribution approximation in (5) is also the basis of useful bounds. In fact, Kingman's bound (Kingman, 1970) is closely related to (5). Kingman proves that setting $\gamma = 1$ in (5) gives a bound, i.e.:

$$\mathrm{P}\{W_q > w\} \leq e^{-\delta w} \tag{8}$$

Kelly (1991) presents a useful interpretation of Kingman's bound for multi-class queues. Consider a queueing system with I classes of customers where the customers of class i arrive according to a Poisson process with rate λ_i and have service time requirements S_i. As in the previous subsections, the resulting system may be analyzed as an M/G/1 queue with arrival rate $\lambda = \sum_{i=1}^{I} \lambda_i$ and service time distribution $S(x)$ where

$$S(x) = \sum_{i=1}^{I} p_i S_i(x)$$

with $p_i = \lambda_i/\lambda$.

Now consider a constraint of the form:

$$P\{W_q > w\} \leq e^{-\gamma} \tag{9}$$

From (8) it is apparent that the bound is satisfied when $\delta w \geq \gamma$ which implies:

$$\lambda \int_0^\infty e^{\gamma x/w}(1 - S(x))dx \leq 1$$

or equivalently

$$\sum_{i=1}^{I} \lambda_i \int_0^\infty e^{\gamma x/w}(1 - S_i(x))dx \leq 1.$$

Letting:

$$\alpha_i := \lambda_i \int_0^\infty e^{\gamma x/w}(1 - S_i(x))dx \tag{10}$$

This can be summarized by the condition:

$$\sum_{i=1}^{I} \alpha_i \leq 1 \tag{11}$$

where α_i is known as the *effective bandwidth* of customer type i and in this case can be considered as a measure of the amount of workload a type i customer brings to the system with respect to the performance constraint (9).

Note that α_i as given in (10) is not dependent on the total arrival rate at the server. This property will prove to be very useful in developing assignment models for multiple servers. Cohen (1994) provides an alternative expression leading to a tighter bound. Unfortunately, in this case α_i's are dependent on the total arrival rate at the server which disables their utility in server assignment type models.

Kelly (1991) proves the existence of similar linear constraints (effective bandwidths) for other performance measures as well. Of particular interest to us is a bound on the average delay. Kelly shows that if the constraint:

$$\mathrm{E}[W] \leq \bar{W} \tag{12}$$

is satisfied, then there exist parameters β_i where:

$$\beta_i = \lambda_i \left[\mathrm{E}[S_i] + \frac{1}{2\bar{W}} (\mathrm{E}[S_i]^2 + \mathrm{Var}[S_i]) \right] \tag{13}$$

such that the performance constraint can be written as the linear constraint:

$$\mathrm{E}[W] \leq \bar{W} \Leftrightarrow \sum_{i=1}^{I} \beta_i \leq 1. \tag{14}$$

In the next section, we will utilize the linear constraints (11) and (14) to formulate account representative assignments under various criteria. But even before that, to motivate the utility of the α_i's, consider the situation in which I sites are served by a single representative designed to give a certain service level guarantee for a response time limit of w. Assume also that when the response time limit is exceeded, a certain penalty is paid. To rank the clients in terms of the penalty cost that they cause, one can simply rank the α_i's (the most costly client is the one with the highest α_i value, since that client consumes the highest proportion of the resource). This way one can determine a standardized cost for each client that depends on the service request rates, repair and transportation times and the service level, thus providing a tool that may assist in pricing decisions.

3 Assigning Multiple Account Representatives to Different Sites

In this section, we consider the model of a field service providing firm that utilizes strict account representative-client assignments as a strategy. Under this strategy each client has its own account representative who attends to all the service requests from this client. Two of the main concerns for the managers

of the firm are probability of exceeding the response time limit as specified in the contract and average response times. Below, we present various models for optimal assignment of representatives to sites under the condition that the response time guarantees will be met with a certain probability and that the average response time does not exceed a certain desired level. Our main purpose in this section is to demonstrate how service level guarantees can be handled in a variety of problem settings. We do not provide solution methodologies for particular formulations since the appropriateness of a formulation depends on the situation. Instead, we give references that include solution methodologies when available and we stay within the framework of linear integer programming formulations, for which at least small sized problems can easily be solved by standard software.

Consider a region with I sites, repair requests from each location arrive according to a Poisson process with rate λ_i. We assume that when a representative j is assigned to a location i each repair takes a random amount of time S_{ij} with distribution $S_{ij}(x)$. Note that the service time distribution allows technicians to have different base to location distances if the round-trips to base type travel model is considered.

Initially, assume that the firm gives identical delay limits, w, to each client as well as requiring that the delay limits are met with probability determined by a parameter γ; i.e., it is required that $P\{W_i > w\} \leq e^{-\gamma}$ where W_i is the random variable that denotes the waiting time in the ith site. As in the previous section, we can define an effective bandwidth, α_{ij} for the assignment of representative j to client site i:

$$\alpha_{ij} = \lambda_i \int_0^\infty e^{\gamma x/w}(1 - S_{ij}(x))dx \tag{15}$$

To gain insight into the meaning of α_{ij}, once again consider that the delay constraint is a constraint on the utilization of a resource, where the resource in this case is an account representative that should provide a certain service level. Then, α_{ij} can be interpreted as the proportion of resource j that is consumed by site i (with respect to the allowable utilization level).

Management may also consider upper bounds on the average delay that will be experienced for all of the clients. For an upper bound of $\bar{W}$ we can introduce

the following bandwidth:

$$\beta_{ij} = \lambda_i \left[\mathrm{E}[S_{ij}] + \frac{1}{2\bar{W}} (\mathrm{E}[S_{ij}]^2 + \mathrm{Var}[S_{ij}]) \right] \tag{16}$$

To introduce a mathematical formulation of the assignment problem consider the following decision variables:

$$a_{ij} = \begin{cases} 1 & \text{if representative } j \text{ is assigned to site } i \\ 0 & \text{otherwise} \end{cases}$$

for $1 \leq i \leq I$ and $1 \leq j \leq J$.

The first problem that will be formulated is that of a minimum cost assignment. Assume J representatives are available and assigning the jth representative to the ith site has an associated cost of c_{ij}. This cost may include preferences with respect to each assignment, which may depend on locations of the residences of representatives, past client interactions and so on.

The following integer program finds the minimum cost assignment while satisfying the service level requests:

$$\begin{array}{lcll} \text{Min} \ \sum_{i=1}^{I} \sum_{j=1}^{J} c_{ij}\ a_{ij} & & & \\ \quad \text{s.t} & & & \\ \sum_{i=1}^{I} \alpha_{ij} a_{ij} & \leq & 1 & \text{for } \ j = 1, 2, ...J \\ \sum_{i=1}^{I} \beta_{ij} a_{ij} & \leq & 1 & \text{for } \ j = 1, 2, ...J \\ \sum_{j=1}^{J} a_{ij} & = & 1 & \text{for } \ i = 1, 2, ..I \\ a_{ij} & = & 0 \text{ or } 1 & \text{for } \ i = 1, 2, ..., I, \ \ j = 1, 2, ..., J \end{array}$$

Note that the above formulation is a multiple constraint general assignment problem (studied in depth by Gavish and Pirkul, 1991).

It is not uncommon that multiple response time limits are specified in service time contracts. For example, a contract may specify that 90 % of all service requests will be attended to within 4 hours and 99% of all service requests will be attended to within 12 hours. This is easy to handle in the above formulation by defining new bandwidths using the new delay limit in (15) and adding a corresponding constraint.

Next, we consider a class of staffing problems in which the objective is to find the optimal size of a representative crew under the condition that the

service level constraints are met. This time J is the maximum number of representatives that the firm would be willing to use. Let z_j $(j = 1, 2, ..., J)$ be one if representative j is utilized and zero otherwise. The problem can now be formulated as follows:

$$\begin{array}{lcll} \text{Min} & \sum_{j=1}^{J} z_j & & \\ & \text{s.t} & & \\ \sum_{i=1}^{I} \alpha_{ij} a_{ij} & \leq & 1 & \text{for } j = 1, 2, ...J \\ \sum_{i=1}^{I} \beta_{ij} a_{ij} & \leq & 1 & \text{for } j = 1, 2, ...J \\ \sum_{j=1}^{J} a_{ij} & = & 1 & \text{for } i = 1, 2, ..I \\ z_j & \geq & a_{ij} & \text{for } i = 1, 2, ..., I, \ \ j = 1, 2..., J \\ z_j & = & 0 \text{ or } 1 & \text{for } j = 1, 2, ..., J \\ a_{ij} & = & 0 \text{ or } 1 & \text{for } i = 1, 2, ..., I, \ \ j = 1, 2, ..., J \end{array} \tag{17}$$

If each representative has a different cost c_j depending on experience and skill level, one can alternatively formulate a problem to find the minimum cost crew.

Now consider the special case of the staffing problem in a dense geographical region with uniform representatives. In this case, the repair time distributions do not depend on the particular assignment. This implies that the parameters α_{ij} and β_{ij} do not depend on j, i.e: we have for all i:

$$\alpha_{ij} = \alpha_i \ \text{ and } \ \beta_{ij} = \beta_i \ \text{ for } \ j = 1, 2, .., J \tag{18}$$

Under the above assumptions, one can bound (from below) the number, z_*, of representatives required to attain the performance requirement. This can easily be seen as follows. The first two constraints of the above problem under the new assumptions:

$$\begin{array}{lcll} \sum_{i=1}^{I} \alpha_i a_{ij} & \leq & 1 & \text{for } j = 1, 2, ...J \\ \sum_{i=1}^{I} \beta_i a_{ij} & \leq & 1 & \text{for } j = 1, 2, ...J \end{array} \tag{19}$$

can equivalently be expressed as:

$$\begin{array}{lcll} \sum_{i=1}^{I} \alpha_i a_{ij} & \leq & z_j & \text{for } j = 1, 2, ...J \\ \sum_{i=1}^{I} \beta_i a_{ij} & \leq & z_j & \text{for } j = 1, 2, ...J \end{array} \tag{20}$$

Now summing up both equations over j, exchanging summations on the left hand side of the inequality and using $\sum_{j=1}^{J} a_{ij} = 1$ for all i $(i = 1, 2, .., I)$ we obtain the bound as:

$$z_* \geq \max \left\{ \lceil \sum_{j=1}^{J} \beta_j \rceil, \lceil \sum_{j=1}^{J} \alpha_j \rceil \right\} \tag{21}$$

where $\lceil x \rceil$ denotes the smallest integer greater than x.

To demonstrate the utilization of the staffing type formulations, we present the following numerical examples.

Example 1: Consider 7 sites with breakdown rates (per hour) $\lambda_1 = 0.5$, $\lambda_2 = 0.2$, $\lambda_3 = 0.4$, $\lambda_4 = 0.3$, $\lambda_5 = 0.25$, $\lambda_6 = 0.25$, $\lambda_7 = 0.7$. Assume that the firm gives a response time guarantee of 8 working hours and it is desired that the guarantee is met at least 90% of the time and the maximum number of repair people that could be employed is 5. Consider the case where all $S_{i,j}$ $(i = 1, 2, .., 7,\ j = 1, 2, .., 5)$ are exponentially distributed with mean $1/\mu = 1$ hour. (This corresponds to the case where all clients' requests have identical statistical characteristics independent of the particular representative assignment.)

To set the integer programming formulation, the bandwidths have to be computed. Using equation (15), it is found that: $\alpha_1 = 0.70$, $\alpha_2 = 0.28$, $\alpha_3 = 0.56$, $\alpha_4 = 0.42$, $\alpha_5 = 0.35$, $\alpha_6 = 0.35$, $\alpha_7 = 0.28$.

The solution of the integer program yields that 3 servers are necessary to meet the response time requirements in this case. The following assignments are obtained: repair person 1 to sites 1 and 7, repair person 2 to sites 2,5 and 6, and repair person 3 to sites 3 and 4 (this is not a unique optimum). In this case, the lower bound according to (21) $\lceil \sum_{j=1}^{7} \alpha_j \rceil$ is also equal to 3.

We can also measure the actual performance of the system under the above assignment since each repair person operates as an M/M/1 queue for which the response time distribution is:

$$W(x) = 1 - \rho e^{-\mu(1-\rho)x}$$

where $\rho = \lambda/\mu$.

For the above example $\rho_1 = \lambda_1 + \lambda_7 = \rho_2 = \lambda_2 + \lambda_5 + \lambda_6 = \rho_3 = \lambda_3 + \lambda_4 = 0.7$ which implies that the probability of meeting the service guarantee is 0.94.

Example 2: It would be interesting to quantify the costs that are imposed by using the fixed assignment of representatives in comparison to a fully flexible scheme where a pool of technicians is used to deal with incoming service requests. In a fully flexible scheme any technician can be assigned to any client which makes the system a single M/M/c queue for which the waiting time distribution is explicitly available (see Gross and Harris 1985, for example). Consider the following case: there are four different client classes classified according to repair request rates. There are 30 clients of type 1, 20 of type 2, 10 of type 3 and 10 of type 4 with respective service request rates of 0.05, 0.1, 0.2 and 0.3 per hour for each class (where each client's request arrives according to an independent Poisson process). One can then use (21) to obtain the minimum crew size to achieve given service levels. The results are summarized in Figure 2.

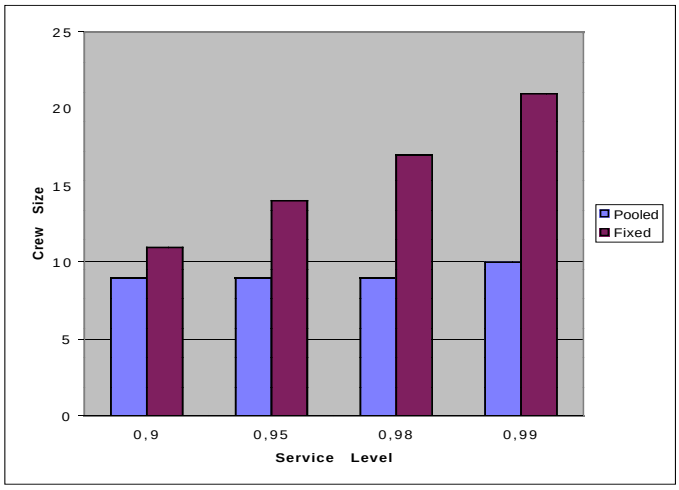

Figure 2: Required Crew Sizes for Different Service Levels

Even these simple examples demonstrate that dedicated service has a considerable cost. On the other hand, it is obvious that complete resource pooling is both hard to implement and will meet with strong opposition from the clients. In section 4 we investigate this issue in further detail to understand the condi-

tions under which dedication is a viable alternative.

In previous formulations a fixed service level was assigned to all customers in the service network. Since service warranties are usually individual contract based, it is also of interest to model the case in which the service levels of different customers can be selected individually depending on cost and revenue considerations.

To remain within the realm of assignment models we have to assume that in this situation each client can be offered only one of k ($k = 1, 2, ..., K$) different classes of service level and target response time combinations. A type k combination is determined by a pair (w_k, γ_k) where w_k specifies the target response time and $e^{-\gamma_k}$ specifies the service level probability such that :

$$P\{W > w_k\} \leq e^{-\gamma_k}$$

This immediately leads to a bandwidth:

$$\alpha_{ijk} = \lambda_i \int_0^\infty e^{\gamma_k x / w_k} (1 - S_{ij}(x)) dx \tag{22}$$

As a shortcoming of this type of formulation, note that, α_{ijk}'s are meaningful only if all clients assigned to a representative have identical target response times and service levels. Hence in the formulation we add the constraint that each representative is assigned a single service level/response time combination.

As a specific instance of this model, consider the case where assigning client i service level k brings an associated revenue of r_{ik} which reflects the direct revenue specified by the contract as well as perhaps future considerations such as client retention.

Define the following decision variables:

$$a_{ijk} = \begin{cases} 1 & \text{if representative } j \text{ is assigned to site } i, \text{ while site } i \text{ is served at level } k \\ 0 & \text{otherwise} \end{cases}$$

for $1 \leq i \leq I$, $1 \leq j \leq J$, $1 \leq k \leq K$, and:

$$z_{jk} = \begin{cases} 1 & \text{if representative } j \text{ is designated as a type } k \text{ service level provider} \\ 0 & \text{otherwise} \end{cases}$$

for $1 \leq j \leq J$ and $1 \leq k \leq K$.

$$\text{Max} \quad \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} r_{ik} a_{ijk}$$

s.t

$$\begin{aligned}
\sum_{k=1}^{K} \sum_{i=1}^{I} \alpha_{ijk} a_{ijk} &\leq 1 && \text{for } j = 1, 2, ...J \\
\sum_{j=1}^{J} \sum_{k=1}^{K} a_{ijk} &= 1 && \text{for } i = 1, 2, ..I \\
z_{jk} &\geq a_{ijk} && \text{for } i = 1, 2, ..., I, \ j = 1, 2..., J, \\
& && k = 1, 2, .., K \\
\sum_{k=1}^{K} z_{jk} &= 1 && \text{for } j = 1, 2, ..., J \\
z_{jk} &= 0 \text{ or } 1 && \text{for } j = 1, 2, ..., J, \ k = 1, 2, .., K \\
a_{ijk} &= 0 \text{ or } 1 && \text{for } i = 1, 2, ..., I, \ j = 1, 2, ..., J, \\
& && k = 1, 2, ..., K
\end{aligned}$$

4 Almost Fixed Assignments

In the previous section, we discussed performance and optimization issues when each client is assigned to a unique account representative who handles all service requests. Simple numerical examples have shown that additional cost of dedication over complete pooling can be very high. On the other hand, it is evident that clients' requests for dedication eliminate the possibility of complete resource pooling. Therefore, a compromise may be necessary. In this section, we investigate the effects of including some degree of flexibility in the fixed account representative system. In particular, we are interested in measuring the performance of a system in which account representatives are responsible to handle most (but not all) requests from their clients. Our purpose here is not to provide a detailed scheduling analysis but to extract guidelines for design. Hence, we introduce a simple model that nevertheless captures the essence of various tradeoffs between fixed client-representative assignments and more flexible ones.

4.1 Dynamic Assignment of Service Requests with Flexibility Constraints

Consider two representatives that each have their unique patch of clients. Under a fixed assignment scheme, the two servers form two parallel queues each with the customer requests forming the arrival streams. Let λ_1 (λ_2) denote arrival rate of customers that are in the patch of representative 1 (2). Furthermore, let μ_1 and μ_2 be the average service rates of representatives 1 and 2. For simplicity, we also make the assumption that arrivals occur according to Poisson processes and service times are exponentially distributed. One way to introduce flexibility in this system, is by routing a customer that is in the patch of representative 1 (a type 1 customer) towards representative 2 or vice versa. Moreover, if we assume that once a client is assigned, there will be no jockeying between queues, we can measure the number of mismatched customers by counting the number of customers that are assigned to the *not-preferred representative*. We can then quantify the flexibility of the service system by the following ratio:

$$\text{Flexibility Ratio} = \frac{\text{Expected number of mismatches per unit time}}{\text{Expected number of arrivals per unit time}} \tag{23}$$

This way zero flexibility corresponds to a fixed assignment scheme, but by constraining the flexibility ratio we can also obtain almost fixed assignment schemes. However, note that system performance with a flexibility ratio constraint depends on the particular assignment policy used. Therefore, to compare the performance of the system under different flexibility constraints one should be consistent in the assignment policy used. To this end, we will formulate a dynamic control problem that can be solved to yield the optimal dynamic routing policy for a given flexibility ratio constraint. This will allow us to make fair comparisons between different flexibility levels.

Previously, we had underlined two performance measures that are critical for field service systems. The average response time and the probability of exceeding a given response time limit. Below, we formulate a stochastic dynamic program that minimizes the total expected queue lengths (strongly related to average response times) but also evaluate the probabilities of exceeding a given response time limit for the optimal policy.

Let t_n denote the time of the n'th event (arrival or departure from the system), let $Q_j(t_n)$ denote the number of clients waiting to be served in queue j ($j = 1, 2$) (including the customer being serviced) at time t_n and let $I_i(t_n)$ ($i = 1, 2$) be the indicator function corresponding to arrival types, i.e.$I_i(t_n) = 1$ if the event at time t_n is an arrival of type i and $I_i(t_n) = 0$ otherwise. We can now denote the state of the system by the vector $X_n = (Q_1(t_n^-), Q_2(t_n^-), I_1(t_n), I_2(t_n))$ where the state space is $\mathcal{N} \times \mathcal{N} \times \{0, 1\} \times \{0, 1\}$.

The controls correspond to the routing decisions and they depend on the type of customer that arrives, define $U_i(t_n)$ ($i = 1, 2$) as the routing decision to be made at the instance of the n'th event. Besides the queue lengths, we are also interested in the number of *mismatches*. A mismatch will arise when a customer of type i is routed to queue j ($j \neq i$) upon arrival. We set $U_i(t_n) = 1$ to mark a mismatch, hence $U_i(t_n) = 1$ if $I_i(t_n) = 1$ and customer i is routed to queue j ($j \neq i$), otherwise $U_i(t_n) = 0$.

Let $c(X_n)$ denote the immediate cost incurred at decision epoch n, then:

$$c(X_n) = Q_1(t_n^-) + Q_2(t_n^-) \tag{24}$$

The constrained Markov decision problem can be set as:

$$\begin{aligned} &\min \ \ \lim_{T\to\infty} \inf \tfrac{1}{T}\mathrm{E}\left[\textstyle\sum_{n:0\le t_n<T} c(X_n)|X_0 = x\right] \\ &\text{subject to} \\ &\lim_{T\to\infty} \inf \frac{\mathrm{E}\left[\sum_{n:0\le t_n<T} U_1(t_n)+U_2(t_n)|X_0=x\right]}{\mathrm{E}\left[\sum_{n:0\le t_n<T} I_1(t_n)+I_2(t_n)|X_0=x\right]} \le \mathrm{p} \end{aligned} \tag{25}$$

When the state space is truncated, the above problem can be formulated as a linear program by standard results in Markov Decision Processes. The linear programming formulation enables the numerical computation of optimal routing policies. However, it would also be useful to understand the structure of good routing policies to see if they can be approximated by simple rules that are implementable in practice. To this end, we relax the constraint in (25) by adding it to the objective function. Therefore, we now have the unconstrained Lagrangean problem with penalties for each mismatch r:

$$\min \ \lim_{T\to\infty} \inf \frac{1}{T}\mathrm{E}\left[\sum_{n:0\le t_n<T} c(X_n) + rU_1(t_n) + rU_2(t_n)|X_0 = x\right] \tag{26}$$

The relaxed version of the problem is an unconstrained MDP for which the optimal policy is stationary and non-randomized. Moreover as will be shown in the following theorem, the optimal stationary policy has monotonicity properties that can be exploited to design practically useful control policies.

Theorem 1 Monotone controls separated by increasing switching curves exist for the dynamic assignment problem with mismatch penalties.

Proof: See Appendix A.

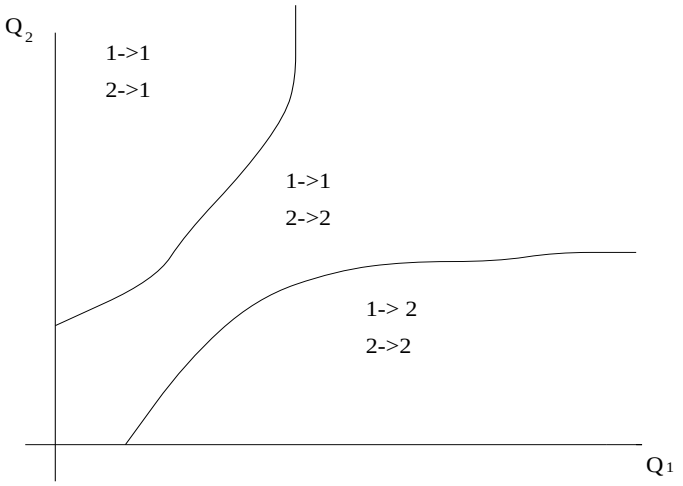

Figure 3: Typical Switching Curves

In Figure 3 we demonstrate a typical optimal assignment policy. In between the two switching curves no mismatch is necessary. Below the first curve (the one closer to the Q_1 axis), the optimal action routes a customer of type 1 to queue 2. Above the second curve the optimal action is to route a type 2 customer to queue 1.

Although exact optimal policies are difficult to implement practically, the monotonicity result enables us to recognize the form of "good" scheduling policies. For example, a class of policies that give piecewise linear monotone switching curves is the so-called (n, N) class of policies introduced by van der Duyn Schouten and Vanneste (1990) under a different setting. As an adaptation of

this class of policies, we propose using two parameters for each switching curve, i.e. (n, N) for curve 1 and (m, M) for curve 2, implying:
i) Transfer a customer from queue 1 to queue 2, if $Q_1(t) > N$ and $Q_2(t) < n$
ii) Transfer a customer from queue 2 to queue 1, if $Q_1(t) < m$ and $Q_2(t) > M$

4.2 Evaluating the Benefits of Flexibility

In this section, we compare the effects of flexibility on account representative-client assignment strategies. We consider performance enhancements in two directions: i) the reduction in the average number of clients waiting to be serviced ii) the proportion of clients who are served later than the given response time limit. Our test case is the two client two representative model introduced above. We set both service rates μ_1 and μ_2 equal to 1 and set the respective arrival rates λ_1 and λ_2 equal to each other. This way, changing the arrival rates is equivalent to changing the traffic load, ρ, for each queue ($\rho = \lambda_1/\mu_1 = \lambda_2/\mu_2$). This corresponds to a case where the clients have identical service limits and the initial design assigns an equivalent load on the representatives.

To measure the performance of the system at different flexibility levels, we use dynamic programming in the following way. For given traffic parameters, we compute the approximately optimal stationary policy using the Lagrangean heuristic described in the previous section. Once the approximately optimal policy is obtained, we use value iteration to compute the expected queue lengths, the number of mismatches per unit time and the proportion of clients who exceed the delay limit under the optimal policy (see Appendix B, for the details of this computation).

In Table 1, we compare the average queue lengths for queues with traffic parameters ranging from 0.1 to 0.9 under three flexibility ratio constraints: 0.05, 0.1 and 0.2. We also report the extreme cases. As the extremely rigid case, we consider the fixed client-representative assignment which leads to two parallel (and noninteracting) M/M/1 queues (the zero flexibility case). As the extremely flexible case (no flexibility ratio constraint), we consider the fully flexible model in which incoming service requests are optimally routed

	Flexibility Ratio Constraint				
ρ	0	0.05	0.1	0.2	1
0.1	0.22	0.22	0.20	0.20	0.20
0.2	0.5	0.47	0.47	0.43	0.43
0.3	0.86	0.77	0.77	0.69	0.69
0.4	1.33	1.22	1.13	1.02	1.01
0.5	2	1.74	1.59	1.59	1.43
0.6	3	2.59	2.27	2.23	2.02
0.7	4.67	3.74	3.47	3.05	2.95
0.8	8	6.05	5.38	4.86	4.73
0.9	18	11.73	10.46	10.46	9.83

Table 1: The expected queue lengths as a function of the traffic load and the flexibility ratio constraint

regardless of representative assignments (which leads to shortest queue routing in this case). A flexibility ratio constraint of one is used to denote the fully flexible case even though for balanced traffic loads the constraint will not be tight with shortest queue routing. Table 2 reports the probability of exceeding a delay limit of 8, for the given combination of traffic load and flexibility ratio constraint.

The results in Table 1 and Table 2 demonstrate that certain setups are unsuitable for strict fixed assignment schemes. Figure 4 displays the improvements in performance (average queue lengths) as a percentage of the total possible improvement as a function of the flexibility ratio constraint. It is apparent that the improvement in performance through flexibility becomes more and more significant at higher traffic loads. In fact, at high traffic loads ($\rho \geq 0.8$) even a small flexibility ratio achieves a performance almost as good as that of the full flexibility. This is consistent with the heavy traffic results of Kelly and Laws (1993) who show that in heavy traffic performance is not very dependent on the particular dynamic routing policy. Tail waiting time probabilities in Table

	Flexibility Ratio Constraint				
ρ	0	0.05	0.1	0.2	1
0.1	0.0001	0.0001	0	0	0
0.2	0.0003	0.0002	0.0002	0.0001	0.0001
0.3	0.0011	0.0004	0.0004	0.0003	0.0003
0.4	0.0033	0.0015	0.0010	0.0008	0.0006
0.5	0.0092	0.0034	0.0023	0.0023	0.0016
0.6	0.0244	0.0107	0.0068	0.0059	0.0044
0.7	0.0635	0.0288	0.0220	0.0166	0.0138
0.8	0.1615	0.0887	0.0662	0.0544	0.0498
0.9	0.4044	0.2626	0.2202	0.2045	0.2055

Table 2: Tail waiting time probabilities as a function of the traffic load and the flexibility ratio constraint

2 confirm this phenomenon even in a sharper manner. For response time limit considerations, fixed assignments are costly with respect to flexible structures even at moderate and low traffic loads.

The above results suggest two different approaches for field service design. In one approach, the system is designed in such a way that representatives will be at work (traveling or repair) most of the time. In this case, some flexibility is essential, an ideal approach would be to have service teams of representatives serving a number of clients. On the other hand, even under this design, it is possible to designate a single representative that deals with most (say 90%) of requests from a given client.

The second approach to field service system design is to have representatives working on clients' requests (travel+repair) about half of the time. In this case, the performance improvement attained by flexibility may be less significant than longer term goals of customer retention and high quality service reputation thus permitting a fixed representative assignment structure. One can envision many other longer term benefits in this case. For instance, since the representatives

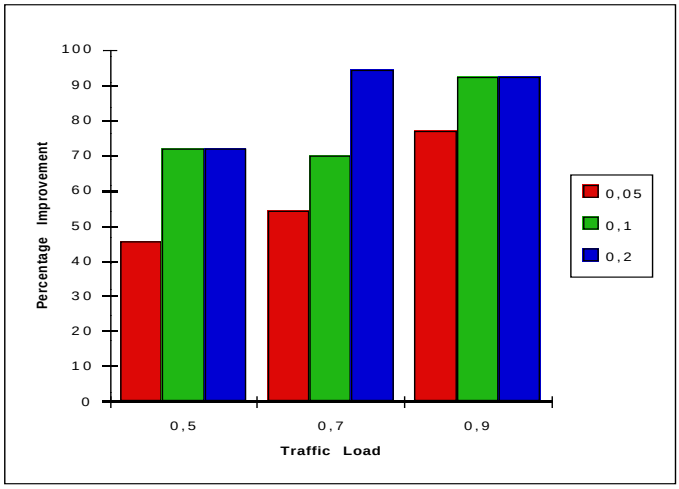

Figure 4: Improvement in Performance (Expected Queue Lengths) at Different Traffic Loads

are not overwhelmed with work, they could work with the clients on a more customized basis. This is a critical issue for services, as the definition of *quality of service* is a function of customer perception and hence is very difficult to standardize. Another important longer term benefit of fixed assignments is that it leaves time and motivation to the representatives to cooperate on new service and product development.

5 Conclusions and Future Research

After sales service becomes increasingly important to many firms in terms of revenues. In addition, the quality of service affects the competitive position of the firm. Studies in the automobile sector, for instance, have shown that better field service leads to increased sales. However, in spite of the increasing importance of field service, it appears that very little research work has been done to date.

We provided planning models for a field service firm striving to combine short term service performance goals with longer term goals of client retention and satisfaction. Based on a particular example of a field service firm, we

recognized the unique customer representative assignment as a prior condition to achieve the longer term goals. We then demonstrated that one can formulate tractable mathematical models that combine service level guarantees with the dedicated service scheme. Finally, we introduced a model that investigates the dedication versus short term performance trade-offs.

It was seen that dedication strongly conflicts with short term performance goals under certain system configurations. In particular, our model demonstrated that if service representatives are overwhelmed with clients' requests most of the time, the strict dedication framework is too costly in the short term. Dedication is preferable when service representatives have looser schedules. We noted that, although expensive in the short run, this may have major advantages for new service/product development and customization.

Many other interesting and important issues for field services cannot be covered by the models in this paper. For example, we noted the high operational cost of a dedicated account representative scheme versus flexible service schemes. One important direction for future research is capturing explicitly the effects of service customization and the account representative framework on contract renewal and market extension. This would permit a better understanding of the trade-offs between longer term benefits versus the short term costs.

One can expect that customization would give field service firms richer pricing options in terms of different quality of service levels. We could provide a partial guideline for the design of a service system under the dedicated service scheme with multiple quality of service levels. In particular, a complete analysis in this direction should investigate the effects of assigning clients with different service quality levels to one and the same service representative. This also brings along the issue of detailed scheduling; a challenging task when clients pay different amounts for services and may have different requirements.

Concluding, we would like to note that a company's competence in field service extends beyond the life cycle of a typical product. Surely, better service may field more sales as well as more revenues form after-sales activities. However, dedicated field engineers also allow for better information on how

the product is used by the customer. This, in turn, may have several spin-offs. First, when the product returns at the end of its useful life or when the lease expires, the producer will have valuable information on the condition of the product and hence on the best recovery option. Second, close attention to a customer's needs may increase customer loyalty and allow for economies of scope through cross-selling. Third, dedicated field engineers can and should be used as valuable sources for new product ideas and/or improvements to existing product lines. Incorporating the above ideas into workable models is an important challenge for future research.

References

Abate J., G.L. Choudhury and W. Whitt, "Exponential Approximations for Tail Probabilities in Queues, 1: Waiting Times", *Operations Research*, Vol. 43, 1995, pp. 885-901.

Agnihothri, S.R., S. Narasimhan and H. Pirkul, "An Assignment problem with Queueing Time Cost", *Naval Research Logistics Quarterly*, Vol. 37, 1990, pp. 231-243.

Amiri, A., "Solution Procedures for the Service System Design Problem", *Computers and Operations Research*, Vol. 24, 1997, pp. 49-60.

Bertsimas, D.J. and G. van Ryzin, "Stochastic and Dynamic Vehicle Routing in the Euclidean Plane with Multiple Capacitated Vehicles", *Operations Research*, Vol. 41, 1993, pp. 60-76.

Cohen, J.W., "On the Effective Bandwidth Design for the Multi-Server Channels", *Research Report BS-R9406, CWI, Amsterdam*, 1994.

Cohen, M.A., Y.-S. Zheng and V. Agrawal, "Service Parts Logistics: A Benchmark Analysis", *IIE Transactions*, Vol. 29, 1997, pp. 627-639.

Gavish, B. and H. Pirkul, "Algorithms for the Multi Resource Generalized Assignment Problem", *Management Science*, Vol. 37, 1991, pp. 695-713.

Gross, D. and C.M. Harris, *Fundamentals of Queueing Theory, Second Edition*, John Wiley and Sons, New York, 1985.

Hambleton, R.S., "A Manpower Planning Model for Mobile Repairman", *Journal of the Operational Research Society*, Vol. 33, 1982, pp. 621-627.

Hajek, B., "Optimal Control of Two Interacting Service Stations", *IEEE Transactions on Automatic Control*, Vol. 29, 1984, pp. 491-499.

Hill, A.V., S.T. March, C.J. Nachtsheim and M.S. Shanker, "An Approximate Model for Field Service Territory Planning", *IIE Transactions*, Vol. 24, 1992a, pp. 2-10.

Hill, A.V., "An Experimental Comparison of Dispatching Rules for Field Service Support", *Decision Sciences*, Vol. 23, 1992b, pp. 235-249.

Kelly, F.P., "Effective Bandwidths at Multi-class Queues", *Queueing Systems*, Vol. 9, 1991, pp.5-16.

Kelly, F.P. and C.N. Laws, "Dynamic Routing in Open Queueing Networks: Brownian Models, Cut Constraints and Resource Pooling", *Queueing Systems*, Vol. 13, 1993, pp. 47-86.

Kingman, J.F.C., "Inequalities in the Theory of Queues", *Journal of the Royal Statistical Society*, Series B32, 1970, pp. 102-110.

Koole, G., "Structural Results for the Control of Queueing Systems Using Event-Based Dynamic Programming", *Working Paper WS-461, Faculty of Mathematics and Computer Science, Free University, Amsterdam*, 1996.

Melachrinoudis, E., "A Discrete Location Assignment Problem with Congestion", *IIE Transactions*, Vol. 26, 1994, pp. 83-89.

Smith, S.A., "Estimating Service Territory Size", *Management Science*, Vol. 25, 1979, pp. 301-311.

Tijms, H. , *Stochastic Models an Algorithmic Approach*, John Wiley and Sons, New York, 1994.

Van der Duyn Schouten, F.A. and S.G. Vanneste, "Analysis and Computation of (n,N) Strategies for Maintenance of a Two-component System", *European Journal of Operational Research*, 1990, Vol. 48, pp. 260-274.

Weber R.R. and S. Stidham, "Optimal Control of Service Rates in Networks of Queues", *Advances in Applied Probability*, Vol. 19, 1987, pp. 202-218.

Appendix A

Proof of theorem:

We first consider the discounted case of the problem, with a discount factor

of α, i.e:

$$\min \lim_{T->\infty} \inf \mathrm{E}\left[\sum_{n:0\leq t_n<T} e^{-\alpha t_n}(c(X_n)+rU_1(t_n)+rU_2(t_n))|X_0=x\right] \quad (27)$$

We uniformize the transition rates such that:

$$\phi = \lambda_1+\lambda_2+\mu_1+\mu_2+\alpha \quad (28)$$

and choose $\phi = 1$ without loss of generality.

The corresponding optimality equations read as follows:

$$\begin{aligned} V_\alpha(x_1,x_2) &= c(x_1,x_2)+\lambda_1 \min\{V_\alpha(x_1+1,x_2), r+V_\alpha(x_1,x_2+1)\} \\ &\quad \lambda_2 \min\{r+V_\alpha(x_1+1,x_2), V_\alpha(x_1,x_2+1)\} \\ &\quad +\mu_1 V_\alpha((x_1-1)^+,x_2)+\mu_2 V_\alpha(x_1,(x_2-1)^+) \end{aligned} \quad (29)$$

where $(x)^+$ denotes $\max\{0,x\}$.

Let $V_\alpha^k(x_1,x_2)$ denote the k stage value function, to prove that increasing switching curves exist, we now define the following difference functions:

$$\Delta_{1,\alpha}^k(x_1,x_2) = V_\alpha^k(x_1+1,x_2)-V_\alpha^k(x_1,x_2+1) \quad (30)$$

$$\Delta_{2,\alpha}^k(x_1,x_2) = V_\alpha^k(x_1,x_2+1)-V_\alpha^k(x_1+1,x_2) \quad (31)$$

The following lemma implies the existence of increasing switching curves for the discounted cost case.

Lemma 1 *$\Delta_{1,\alpha}(x_1,x_2)$ is monotonically nondecreasing in x_1 for fixed x_2 and $\Delta_{2,\alpha}(x_1,x_2)$ is monotonically nondecreasing in x_2 for fixed x_1.*

Proof: Using the fact that $V_\alpha = \lim_{k\to\infty} V_\alpha^k$, we argue inductively by value iteration starting with $V_\alpha^0(x_1,x_2)=0$. We need to show that the desired property propagates through value iteration. It turns out that the propagation of the desired property at the boundaries of the state space (i.e. when x_1 or x_2 equals 0) imposes two additional conditions on the function V_α^k. The first additional condition is that $V_\alpha^k(x_1,x_2)$ is increasing in x_1 for fixed x_2 and increasing in x_2 for fixed x_1. The second condition requires supermodularity of V_α^k, i.e.:

$$V_\alpha^k(x_1+1,x_2)+V_\alpha^k(x_1,x_2+1) \leq V_\alpha^k(x_1,x_2)+V_\alpha^k(x_1+1,x_2+1) \quad (32)$$

At this point, we can refer to Lemma 4.1 of Koole (1996) where it is shown that the departure operators, $\mu_1 V_\alpha^k((x_1-1)^+, x_2)$ and $\mu_2 V_\alpha^k(x_1,(x_2-1)^+)$, as well as the controlled arrival operators, $\lambda_1 \min\{V_\alpha(x_1+1,x_2), r+V_\alpha(x_1,x_2+1)\}$ and $\lambda_2 \min\{r+V_\alpha(x_1+1,x_2), V_\alpha(x_1,x_2+1)\}$ propagate the necessary properties. Since the linear holding cost function $c(x_1,x_2)$ also propogates these properties, the result follows.

□

Finally, to complete the proof it is required to show that the monotonicity property shown in Lemma 1 for the discounted case carries over the average cost case expressed in (26). Let $\Delta_i(x_1,x_2)$ (i=1,2) be the undiscounted version of $\Delta_{i,\alpha}^k(x_1,x_2)$.

Lemma 2 *$\Delta_1(x_1,x_2)$ is monotonically nondecreasing in x_1 for fixed x_2 and $\Delta_2(x_1,x_2)$ is monotonically nondecreasing in x_2 for fixed x_1.*

Proof: We first argue that the average cost problem can be treated as the limit of discounted case problems since the conditions in Weber and Stidham (1987) are verified. The relative difference function for the average cost case can be expressed as:

$$V(x_1,x_2) = V_\alpha(x_1,x_2) - V_\alpha(0,0) \tag{33}$$

as the discount rate α tends to zero. It is apparent then, that $V(x_1,x_2)$ also satisfies all of the conditions that have been mentioned in the proof of Lemma 1. Therefore the monotonocity of Δ_1 and Δ_2 follow.

□

Appendix B: Let $\mathcal{P}$ denote the policy that is obtained through the Lagrangean heuristic and let $U_i(x_1,x_2,\mathcal{P})$ denote the (stationary action) corresponding to a type i customer arrival under policy $\mathcal{P}$ with x_i customers of type i in the system. Note that, $U_i(x_1,x_2,\mathcal{P})=1$ denotes a mismatch.

B1. Expected Queue Lengths

To compute the expected queue lengths under the policy $\mathcal{P}$, we apply the value iteration technique for semi-Markov processes and define the operator, $\mathcal{T}_{\mathcal{P}}$, given by:

$$\begin{aligned}\mathcal{T}_{\mathcal{P}}V^k(x_1,x_2) &= x_1+x_2+\lambda_1(1-U_1(x_1,x_2,\mathcal{P})))V^k(x_1+1,x_2)\\ &\quad +U_1(x_1,x_2,\mathcal{P})V^k(x_1,x_2+1))\\ &\quad +\lambda_2(1-U_2(x_1,x_2,\mathcal{P}))V^k(x_1+1,x_2)\\ &\quad +U_2(x_1,x_2,\mathcal{P})V^k(x_1,x_2+1))\\ &\quad +\mu_1V^k((x_1-1)^+,x_2)+\mu_2V^k(x_1,(x_2-1)^+)\end{aligned} \tag{34}$$

The expected total queue length L_1+L_2 can then be obtained by using the relation:

$$V^{k+1}(x_1,x_2)+(L_1+L_2)=\mathcal{T}_{\mathcal{P}}V^k(x_1,x_2) \tag{35}$$

and letting $k\to\infty$ with $V^0=0$.

B2. Tail Probabilities

Let $w_j(x_j)$ denote the probability that the waiting time of a client assigned to queue j will exceed the response time limit when there are x_j customers waiting to be served and let A_d be the number of arrivals per unit time that exceed the limit.

Define the operator $\mathcal{T}'$

$$\begin{aligned}\mathcal{T}'_{\mathcal{P}}V^k(x_1,x_2) &= \lambda_1(1-U_1(x_1,x_2,\mathcal{P}))(w_1(x_1+1)+V^k(x_1+1,x_2))\\ &\quad +U_1(x_1,x_2,\mathcal{P})(w_2(x_2+1)V^k(x_1,x_2+1))\\ &\quad +\lambda_2(1-U_2(x_1,x_2,\mathcal{P}))(w_1(x_1+1)+V^k(x_1+1,x_2))\\ &\quad +U_2(x_1,x_2,\mathcal{P})(w_2(x_2+1)+V^k(x_1,x_2+1)))\\ &\quad +\mu_1V^k((x_1-1)^+,x_2)+\mu_2V^k(x_1,(x_2-1)^+)\end{aligned} \tag{36}$$

Hence, starting from $V(0)=0$, A_d can be obtained from the recursive relation:

$$V^{k+1}(x_1,x_2)+A_d=\mathcal{T}'_{\mathcal{P}}V^k(x_1,x_2) \tag{37}$$

and letting $k\to\infty$.

The probability of exceeding the delay limit is then given by: $A_d/(\lambda_1+\lambda_2)$