DESIGNING FLEXIBILITY: CHARACTERIZING THE VALUE OF CROSS-TRAINING PRACTICES

O. Zeynep Akşin[*]
and
Fikri Karaesmen[†]

[*] INSEAD
Boulevard de Constance
77300, Fontainebleau Cedex FRANCE

[†] Laboratoire Productique Logistique
Ecole Centrale Paris
Grande Voie des Vignes
92295 Chatenay-Malabry Cedex FRANCE

zeynep.aksin@insead.edu, fikri@pl.ecp.fr

February 2002

Designing Flexibility: Characterizing the Value of Cross-Training Practices

O. Zeynep Akşin [*] Fikri Karaesmen [†]

February 2002

Abstract

Most service systems consist of multi-departmental structures corresponding to multiple types of service requests, with possibly multi-skill agents that can deal with several types of service requests. The design of flexibility in terms of agents' skill sets and assignments of requests is a critical issue for such systems. We explore the questions of how much flexibility to have in terms of the number of skills of the agents and what type of flexibility one would like to have in terms of the composition of agents' skill sets. Our objective is to identify preferred flexibility structures when demand is random and capacity is finite. We start with the analysis of a model where time evolves in discrete time periods and capacities are considered to be deterministic. In this setting, we identify general structural properties of flexibility design pertaining to the marginal values of flexibility and capacity. We define balance in skill diversity and show when structures with higher balance are superior. We then extend the investigation to a setting, where time evolves continuously and each department is a queueing system. Based on a bounding technique from the "dynamic routing" literature, we characterize conditions under which a similar skill diversity balance result holds. Finally, through a numerical study, we illustrate the implications of the structural results in a call center setting.

[*] INSEAD, Boulevard de Constance, 77305 Fontainebleau Cedex, France
[†] Ecole Centrale Paris

1 Introduction

This paper considers service systems with multi-departmental structures having possibly multi-skill servers that treat several types of service requests. Departments represent physical or logical separations of the servers by their primary skills. These primary skills or specializations can be broadened through cross-training. In any such system, it is possible to have a different mix of skill sets with a different number of servers belonging to each skill set. It is well known that more flexibility leads to better operational performance. However given that there are costs associated with creating and maintaining this flexibility, and difficulties managing the resulting more complex system, it is desirable to understand the value of this flexibility in more depth. For an operations manager, the first question that comes to mind is: How many people does one keep as specialists and how many are cross-trained? However, this question cannot be answered before understanding the exact nature of flexibility brought by a cross-trained server. This paper will focus on the latter issue, and try to provide a better understanding of the relationship between different flexibility structures and value. The following questions will be explored: How many skills should servers have (*how much flexibility*)? What are the ideal skill-sets for those that are cross-trained (*what type of flexibility*)? How should these skill-sets be formed in a multi-departmental structure, where each server has a primary skill and some secondary skills (*where*)? These set of questions motivate our research and will be labeled as the flexibility design problem in the ensuing analysis. A modeling framework is proposed that can be used to determine the value of a given flexibility design from an operational standpoint, and to begin to answer some of these questions.

An important application of the flexibility design problem is in the evaluation of cross-training practices for call centers, particularly multilingual call centers. In Europe, many large corporations operate multilingual call centers. Ireland is rapidly becoming a hub for these centers, with companies like Gateway, Dell, Intel, UPS, etc. locating their call centers there, providing service in various European languages (Turek, 2000). For example, Compaq's call center can handle calls in eleven different languages. In addition to language skills, service representatives in these centers need to understand cultural differences between callers, and to be proficient in the technical aspects of the calls that they deal with. As different calls come in, different skills will be required. A call can be handled by the service representatives that have the matching skills. It is clear that in this type of an environment, overlapping skill sets, i.e. flexible service representatives that can deal with multiple languages or multiple types of calls, are highly desirable from an operational standpoint. It is essential to understand the value of different skill sets, and to build a call center that has the right level of flexibility, in the right places, in order to achieve a low cost high service level operation. With staffing costs constituting up to seventy percent of total costs (Koole and Mandelbaum, 2001), the skill set design problem is of critical importance to call centers.

To manage incoming calls to a multilingual call center, or any other call center where agents have multiple skills that match the needs of different calls, routing software is used that assigns calls to agents. Primary and secondary skills of the servers are captured using priorities in these systems. Today, skills-based routing technology has emerged as a growing business in this segment, to address the desire of call center managers to match call needs as closely as possible to agent skills. While the cost of a skills-based routing system is known, the value or benefit of such a system depends on how it is used. Different skill sets in the center will imply different performance effects. Similarly, different ways of routing calls will imply differences in performance. Managers need to design the ideal skill sets for their center. In doing this, questions like which skill sets need to be developed, which ones need more people, and which ones should overlap have to be answered in light of the different routing options. In practice, the adoption of skills-based routing functionality in ACD (automatic call distribution) or CTI (computer telephony integration) systems has been low, partially due to the difficulty call center managers have in assessing the value of such functionality (Sulkin, 2000; Dawson, 1997). To assess the value of such systems, the approach developed in this paper allows the comparison of the value of different skill set designs under any routing scheme. Thus it enables the quantification of the value of a skills-based routing system for a given flexibility design.

A call center with servers organized by their primary skills, or any other service system with a multi-departmental structure can be modeled as multi-class, multi-server queuing systems. For such systems, to assess the value of a given flexibility design, or to compare different designs, one needs to ensure that customers are routed or assigned to the available servers in a dynamically optimal fashion. Under some non-optimal routing scheme the comparison will either be meaningless or only applicable under that particular type of routing policy. The difficulty of the dynamic routing problem for multi-class, multi-server queues is well established. While the stochastic dynamic nature of the problem poses an important methodological challenge, ignoring this characteristic would misrepresent the value of flexibility which is realized precisely in these kinds of settings. We look at this problem in two different ways. The first of these takes a simplified view, where demand is uncertain, but capacity in each department is assumed to be deterministic. Time is considered to evolve in discrete time periods, and unmet demand in any period is lost, implying that time periods are not linked. The analysis in this setting provides the basis of the structural issues that are explored in the subsequent setting, where the service system described above is modeled as a stochastic network with dynamic routing of calls. Our modeling approach for this setting builds on the literature on dynamic routing in networks Laws (1992), Kelly (1994), Bertsimas and Chryssikou (1999). As in the latter two, rather than determining the optimal dynamic routing policy, an upper bound for the value of a given flexibility design under any routing scheme will be used as a metric for the comparisons.

The remaining parts of this paper are organized as follows. Related literature is reviewed in the next section. In Section 3, the service systems being studied are formally introduced and analyzed for the case when demand is uncertain, capacity is deterministic, and time evolves in discrete time periods. Structural results that relate flexibility and performance in this setting are established. In Section 4, each department is modeled as a queue, and time is taken to be continuous. This yields the stochastic dynamic version of the deterministic capacity service system from Section 3. An upper bound on the performance of a given flexibility design for these systems is proposed. Some characteristics of this bound are developed and results from the deterministic capacity, discrete time setting are extended. Section 4.2 takes this general framework and illustrates how one would apply it in the context of a specific application, that of a call center in this case. Section 5 presents a numerical study that develops managerial guidelines for cross-training practice design in call centers, using the insights gained from the analysis of the model. The paper ends with conclusions and some directions for future research.

2 Literature Review

The importance of flexibility in service delivery is well known. A significant source of service delivery process flexibility comes from the use of cross-trained servers. While the practice itself is widespread, there is little formal evaluation of the value of this type of practice from an operations standpoint. In service delivery, very few studies have explored the problem of designing cross-training practices (Pinker and Shumsky, 2000). Numerous studies in manufacturing have looked at the case of flexible workers and their impact on performance in terms of operational measures like throughput. Most of these studies analyze specific work-sharing schemes in queueing network models. (Van Oyen et al., 1999, and references therein) Van Oyen et al. develop models of serial production systems to determine the value of cross-training and to characterize the factors that affect the opportunity for workforce flexibility. These papers assess the value of certain workforce flexibility practices in given settings, however do not tackle the broader question of designing the type of flexibility in these systems. The objective of this paper is to develop a basic framework that will enable a systematic look at these design questions.

More generally, the benefits and design of flexibility in operations have been studied extensively (DeGroote, 1994; Sethi and Sethi, 1990). An important stream consists of papers that address the capacity investment problem in the presence of flexible resources (Fine and Freund, 1990; Van Mieghem 1998; Netessine et al., 2000). These papers assume a certain form of flexibility and then explore the question of the ideal level of this flexibility and how it relates to value under uncertain demand. In this paper, we consider capacity to be fixed and explore the relationship between different flexibility structures and value, without explicitly addressing the optimal capacity issue. In this regard, our analysis parallels that in Jordan and Graves (1995). Focusing on process flexibility, Jordan and Graves (1995) explore the problem of assigning multiple products to multiple plants, where the flexibility of the plants determines which products they can handle. The authors illustrate that well designed limited flexibility is almost as good as full flexibility. To address the question of where flexibility should be added in a system, the authors define a *chain* structure as a group of directly or indirectly connected group of products and plants. It is shown that a structure that enables the formation of fewer long chains is superior to one with multiple short chains. The principles are illustrated with simulations based on real data and subsequent analytical justifications. Hopp et al. (2001) explore the benefits of chaining in the context of cross-training for production lines. Finally, Gurumurthi and Benjaafar (2001) present a numerical investigation of the benefits of chaining based on a queueing model.

Laws (1992), Kelly and Laws (1993), Harrison and Lopez (1999) use heavy traffic analysis to study the problem of dynamic routing in stochastic networks. The system considered in Harrison and Lopez (1999) is very similar to the one studied in this paper: multiple types of customers or jobs and servers with different capabilities. Through a heavy traffic analysis, the authors show that for systems where the capabilities of the servers overlap in a certain way, defined as a form of *communication*, the system collapses to one with a single state, thus resulting in full pooling. Qualitatively, this result is analogous to the one shown by Jordan and Graves (1995). In particular, one observes that both papers establish the fact that limited flexibility is almost as good as full flexibility (full pooling) when carefully constructed. In Jordan and Graves, limited flexibility in the form of long chains approaches full flexibility performance. In Harrison and Lopez, it is limited flexibility with communicating servers that collapses the system to one of full pooling. Taking either one of the approaches in these two papers, one can reach the conclusion that chains (or communicating servers) are good. However, neither of these papers take the extra step in saying how different chains compare, or what some of the structural properties of superior chain structures may look like. In the heavy traffic environment in Harrison and Lopez, all server configurations that satisfy the communication property will collapse to a single state system, hence will be identical. This approach does not enable a comparison of different systems that satisfy the same communication property. Some guidelines are suggested in Jordan and Graves, where the importance of this issue for managers is emphasized. However these results are not formalized in the analysis. The current paper will revisit the flexibility design problem in a service delivery context, and explore the question of what type of flexibility in further depth. In doing this, we will both investigate the performance effect of different chain structures, and extend the analysis of this problem to a stochastic dynamic setting.

3 Process flexibility in service systems with deterministic capacity

Consider a service system with multiple customer types. Customer types differ in terms of their service requirements. Servers specialize by customer type, but can be flexible with overlapping skill sets, allowing them to treat customer requests from different types. The service system can be represented as a directed graph $G = (N, A)$ with a set of nodes N and a set of arcs A whose elements are ordered pairs of distinct nodes. Some standard definitions are useful to formalize the description of this network. A directed arc (i, j) emanating from node i is said to have tail i, terminating in node j known as the head of the arc. For an arc $(i, j) \in A$, the node j is said to be adjacent to node i. The node adjacency list $A(i)$ is the set of adjacent nodes, $A(i) = \{j \in N : (i, j) \in A\}$. The indegree of a node is the number of incoming arcs of that node and its outdegree is the number of outgoing arcs. For a set of nodes I and J, $(I, J) = \{(i, j) : i \in I, j \in J\}$ represents the set of all arcs between I and J. The capacity of an arc (i, j) is given by the capacity function $c(i, j)$. For subsets I and J of N, denote the sum of all capacities on all arcs (I, J) by $c(I, J) = \sum_{i \in I, j \in J} c(i, j)$. Let X be a subset of N. Then X is a cut if it contains the source but not the sink of the network. The cut capacity function is given by $f(X) = c(X, N \setminus X)$. For X being a cut, the minimum of $f(X)$ over all X is known as a minimum cut.

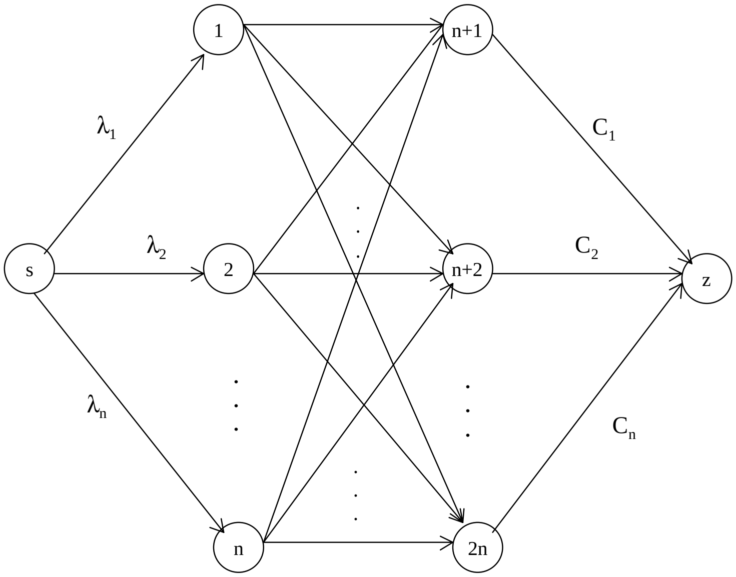

Figure 1: An n class service system with full flexibility

An instance of a network that represents the service system for one time period is depicted in Figure 3. This graph illustrates a system with n customer types given by the set of nodes $I = \{1, 2, ..., n\}$, served by servers in n departments, given by the set of nodes $J = \{n + 1, n + 2, ..., 2n\}$. Note that since servers are assumed to be organized by their primary skills, the number of customer types is equal to the number of departments. The arcs emanating from the source node s and terminating in nodes $i \in I$ represent the service demand, and have capacity given by the demand vector $\lambda = (\lambda_1, ..., \lambda_n)$. This vector represents the realization of demand for a given period. The arcs emanating from nodes $j \in J$ and terminating in the sink node z represent the capacity of each department. These arcs have a capacity given by the vector $\mathbf{C} = (C_1, ..., C_n)$. The arcs (i, j) with $i \in I$ and $j \in J$ represent the flexibility of the system. Whenever a customer of type $i \in I$ can be served by a server of type $j \in J$, an arc (i, j) with infinite capacity is added to the network. The network in Figure 3 illustrates a case where all customers can be treated by all servers, i.e. where the system has full flexibility. In a system with n departments, full flexibility implies that each node $i \in I$ has outdegree equal to n. In general, the outdegree of node $i \in I$ represents the number of possible routings for customers of type i, and the indegree of a node $j \in J$ represents the number of skills a server of type $j - n$ has. Assuming that each customer request of type $i \in I$ is worth r to the system, the problem of maximizing the value generated by a given configuration for a demand realization λ is equivalent to the maximum flow problem for this network.

The optimization problem that we investigate can now be stated formally. It is assumed that the service system evolves in discrete time over the horizon $t = 1, 2, ..T$ (letting $T \rightarrow \infty$, the infinite horizon case can be covered). In the beginning of period t, the demand vector for period t, $\lambda(t) = (\lambda_1(t), \lambda_2(t), ..., \lambda_n(t))$ (where the component $\lambda_i(t)$ denotes the type-i demand in period t), is observed. This demand is then allocated to the departments of the system. Due to capacity constraints, any demand demand that cannot be met with existing capacity in period t is lost. Let $x_i(t)$ denote the number of customers of type-i that are served during period t. The optimal allocation for period t is the allocation that maximizes the period revenue $R(t) = \sum_{i=1}^{n} r_i x_i(t)$. Note that the single period optimal allocation problem is a standard network flow problem.

Let π be the multi-period allocation policy that that uses the optimal single-period allocation (the solution of a network flow problem) in each period and $E_\pi[R] \equiv (1/T)E_\pi[\sum_{t=1}^{T} R(t)]$ be the expected revenue per unit time over the planning horizon of the (myopic) policy π. Remark that, given a particular system structure F and capacity vector C, π is the optimal policy for the multi-period problem because consecutive periods are decoupled due to the lost demand assumption. $E_\pi[R]$ is then the optimal expected revenue per unit time. The objective of the service system is to design the service structure F using its existing capacity C in order to maximize $E_\pi[R]$. Finally, note that in the case where revenues per call are identical for all customer classes ($r_1 = r_2 ... = r_n$), revenue maximization is equivalent to the optimization of throughput. In this case, $E_\pi[T]$ denotes the optimal expected throughput.

In this setting, configuring the skill sets of the servers in order to best utilize existing capacity in the face of random demand is an important issue. Our objective is to understand the ideal structure of this flexibility in order to maximize the value that can be generated from such a service system. The analysis is performed in value terms, and the costs of different configurations are not considered in this study.

In the ensuing analysis, this network flow structure is used to demonstrate some basic properties of flexibility in service systems. We first focus on the single period problem, and show some properties of flexibility structure for given demand. These properties are then explored in the multiple period setting with uncertain demand. The analysis formalizes some well known results, as well as providing new guidelines about skill set (flexibility) design. Throughout this section, the capacity of each department is assumed to be a deterministic quantity. This assumption is relaxed in the following section. All proofs can be found in the Appendix.

Definition 1 *A symmetric network is defined as a network where i) every customer type can be processed by the same number of server types (departments), i.e. each node $i \in I$ has the same outdegree and ii) every department treats the same number of customer types, i.e. every node $j \in J$ has the same indegree. The flexibility of each symmetric network is denoted by F_k with $k = 1, 2, ..., n$, where k indexes the number of server types that a customer type can be served by (which is equal to the outdegree of the nodes $i \in I$), or equivalently the number of customer types that a server type can treat (which is equal to the indegree of the nodes $j \in J$). A higher index k denotes more flexibility, with $k = 1$ representing the case of specialized servers and $k = n$ the case of fully flexible servers. For $1 < k < n$, F_ks are constructed such that the network is fully connected and contains a cycle.*

Throughout, the maximum flow of a network G is denoted by $T_G(\boldsymbol{\lambda}, \mathbf{C})$. For the special case of a symmetric network with flexibility F_k this expression is replaced by $T(F_k, \boldsymbol{\lambda}, \mathbf{C})$. For notational compactness, these are replaced by T_G and $T(F_k)$ whenever there is no information lost.

Theorem 1 *A flexible service system represented by $G(N, A)$ with $n = 3$, demand vector $\lambda = (\lambda_1, \lambda_2, \lambda_3)$ and symmetric capacity vector $\mathbf{C} = (C, C, C)$ has increasing and concave maximum flow $T(F_k, \lambda, \mathbf{C})$ in k with $k = 1, 2, 3$.*

This result establishes the well known property that the performance of a service system with limited flexibility rapidly approaches that of a system with full flexibility. Theorem 1 states one version of this result and provides a proof that makes use of the network flow structure of the problem. While the result is stated for a network with three classes, it is conjectured that it holds for the general case. An approach that generalizes the proof for Theorem 1 would establish the result for a network with n classes.

Corollary 1 *For the system in Theorem 1, in the multiple period setting with random demand, expected average throughput under the optimal allocation policy π, $E_\pi[T(F_k)]$ is increasing and concave in k : $E_\pi[T(F_2)] - E_\pi[T(F_1)] \geq E_\pi[T(F_3)] - E_\pi[T(F_2)]$.*

The result relies on two observations. First note that since Theorem 1 for the single period case was shown to hold for *any demand realization*, the result can be stated for any sample path in the multiple period case. In addition, the assumption that all unmet demand in a given period is lost ensures that the multiple period problem can be viewed as a sum of separable single period problems. Remark that the result requires no particular assumption on the random demand process.

Theorem 1 and Corollary 1 state that for given fixed capacity, an increase in system flexibility leads to an increase in system performance. The returns to flexibility are shown to be diminishing. A different method of improving performance in such a system would be by adding capacity. For example, one may choose to invest in additional capacity rather than investing in additional flexibility. The close interaction between flexibility and capacity are evident. The following result formalizes this relationship, for the single period problem.

Theorem 2 *Consider a symmetric service network with demand vector $\boldsymbol{\lambda} \in \mathcal{R}^n$ and symmetric capacity vector $\mathbf{C} \in \mathcal{R}^n$. If $\sum_{i=1}^n \lambda_i \geq \sum_{i=1}^n C_i$ then $T(F_k, \boldsymbol{\lambda}, \mathbf{C} + \Delta^C) - T(F_k, \boldsymbol{\lambda}, \mathbf{C}) \geq T(F_{k-1}, \boldsymbol{\lambda}, \mathbf{C} + \Delta^C) - T(F_{k-1}, \boldsymbol{\lambda}, \mathbf{C})$ for small enough Δ^C such that one still has $\sum_{i=1}^n \lambda_i \geq \sum_{i=1}^n C_i + \Delta_i^C$.*

The theorem demonstrates that flexibility and capacity are complements up to a certain point parameterized by the demand and capacity vectors. Note that this point represents $T(F_n, \boldsymbol{\lambda}, \mathbf{C})$ since we know that the minimum cut for such a network takes the value $\min(\sum_{i=1}^n \lambda_i, \sum_{i=1}^n C_i)$. Beyond the cutoff point, capacity may act as a substitute to flexibility.

For the multi-period case with random demand, a counterpart of Corrolary 1 cannot be stated except in the special case of 'chronically overloaded' systems where total demand always exceeds total capacity. Informally, for heavily utilized systems, which are also those systems where system flexibility is sought most, one will mostly be in the former region with capacity and flexibility acting as complements. For these types of systems, the result suggests that an additional server is more valuable in the system with superior flexibility. Thus flexibility and capacity should be jointly designed. For systems with lower utilization, this result may be reversed, and an additional person can be worth more in a system with less flexibility. The high flexibility systems are already able to respond to most of their demand, thus the marginal value of capacity is low, becoming zero for the full flexibility system. On the other hand, the low flexibility systems can still improve performance with additional capacity, leading to a positive marginal value for capacity.

So far, we have shown basic features of flexibility that would be useful in answering the *how much flexibility* type of question. Next, the theory of majorization (Marshall and Olkin 1979) is used to explore the type of flexibility, thereby further refining the notion of *smart limited flexibility.*

Definition 2 *For a vector $x \in \mathcal{R}^n$, let $[i]$ denote a permutation of the indices $\{1, 2, ..., n\}$ such that $x_{[1]} \geq x_{[2]} \geq \ldots \geq x_{[n]}$. Then, for x, $y \in \mathcal{R}^n$, x is said to be majorized by y, $x \prec y$, if $\sum_{i=1}^{n} x_{[i]} = \sum_{i=1}^{n} y_{[i]}$ and for all $k = 1, \ldots, n-1$, $\sum_{i=1}^{k} x_{[i]} \leq \sum_{i=1}^{k} y_{[i]}$.*

Let $i_d(J) \in \mathcal{R}^n$ be the vector of indegrees for $j \in J$, and $o_d(I) \in \mathcal{R}^n$ be the vector of outdegrees for $i \in I$. Recall that the former represents a vector with the number of skills of the servers in each department, and the latter the number of possible routings for each customer class. The following definitions are proposed for the graph $G = (N, A)$.

Definition 3 *If $i_d(J)$ has $i_d(n+1) = i_d(n+2) = \ldots = i_d(2n)$, the service system is said to have balanced skill diversity. Symmetrically, if $o_d(I)$ has $o_d(1) = o_d(2) = \ldots = o_d(n)$, the system is said to have balanced routings. For two networks $G = (N, A)$ and $G'(N', A')$, and skill diversity vectors $i_d(J)$ and $i'_d(J')$ (routing vectors $o_d(I)$ and $o'_d(I')$), whenever $i_d(J) \prec i'_d(J')$ ($o_d(I) \prec o'_d(I')$) the system $G(N, A)$ is said to have more balanced skill diversity (more balanced routings) than $G'(N', A')$.*

Theorem 3 *Consider a network $G(N, A)$ with demand vector $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_n)$, symmetric capacity vector $\mathbf{C} = (C, ..., C)$, and routing vector $o_d(I)$. Whenever $G(N, A)$ does not have balanced skill diversity, one can find $G'(N', A')$ with demand vector $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_n)$, symmetric capacity vector $\mathbf{C} = (C, ..., C)$, routing vector $o'_d(I') = o_d(I)$, and $i'_d(J') \prec i_d(J)$ such that $T_{G'} \geq T_G$.*

Corollary 2 *Consider a network $G(N, A)$ with symmetric demand vector $\boldsymbol{\lambda} = (\lambda, ..., \lambda)$, capacity vector $\mathbf{C} = (C_1, ..., C_n)$, and skill diversity vector $i_d(J)$. Whenever $G(N, A)$ does not have balanced routing, one can find $G'(N', A')$ with symmetric demand vector $\boldsymbol{\lambda} = (\lambda, ..., \lambda)$, capacity vector $\mathbf{C} = (C_1, ..., C_n)$, skill diversity vector $i'_d(J') = i_d(J)$, and $o'_d(I') \prec o_d(I)$ such that $T_{G'} \geq T_G$.*

Corollary 3 *The results in Theorem 3 and Corollary 2 continue to hold in the multi period setting with uncertain demand.*

Corollary 3 follows directly from the observations made for Corollary 1 before. It formalizes similar guidelines suggested by Jordan and Graves, that recommend "equalizing the number of plants (measured in total units of capacity) to which each product in the chain is connected" and "equalizing the number of products (measured in total units of expected demand) to which each plant in the chain is directly connected". In the following section, it is explored whether these guidelines also hold in a setting where capacity is no longer deterministic, time is continuous, and one considers the full generality of a stochastic dynamic network.

In order to summarize the findings of this section, both Corollary 1 and Corollary 3 are shown to hold for any demand process. Thus both results continue to apply even in cases where the demand for one service is greater than another, or in the presence of various correlation structures between the demand of different customer types or in time. On the other hand, the qualitative relationship between flexibility and capacity, stated in Theorem 2, depends on the demand realization $\boldsymbol{\lambda}$ directly. In the latter case the single period relationship in 2 does not extend to multiple-periods.

Our representation of the service system as a network, has allowed us to provide proofs for results that have previously not been shown rigorously, and has provided a coherent framework within which one can analyze process flexibility related issues. We note that both the deterministic capacity and discrete time periods assumptions can be thought of as being restrictive and overly stylized for the call center applications that motivated the analysis. As such, the model in this section can be viewed as a crude representation of reality. In the remaining parts of this paper, we explore whether relaxing these assumptions and allowing capacity to be stochastic and time to be continuous has an impact on the results of this section. Our framework allows us to make a natural transition to stochastic dynamic networks in the following section, thus clearly illustrating how these systems compare and contrast.

4 The Performance of a Given Flexibility Design in Stochastic Dynamic Networks

In this section, we relax the assumption that capacity is deterministic, consider continuous time with queueing rather than discrete time periods, and extend the network flow analysis of the previous section to a stochastic dynamic network setting. In Section 4.1, we review a general approach for bounding the optimal performance of such a network under a given flexibility structure and prove a general flexibility design result. Section 4.2 presents a particular call center example that can be analyzed through the approach outlined in 4.1.

4.1 Review of a General Performance Bound and Implications on Flexibility Design

We follow the approach developed by Kelly (1994), which provides an upper bound on the performance of such a network under any dynamic routing scheme. In the current setting, Figure 3 still represents the type of network considered, however in this section $\mathbf{C}$ denotes the number of servers in each department. Actual throughput of each department will be a random variable, that is a function of $\mathbf{C}$, $\boldsymbol{\lambda}$, as well as the routing policy in place.

The analysis is performed at two levels: the network level and the single department level. The single department level consists of a detailed model of the capacity in each department. The underlying model may differ depending on the context: each department can be modeled as a loss system, or as some form of a queueing system. A department $j = 1, \ldots, n$ receives *direct requests* from customers of type j, with rate $\lambda^d = \lambda_j$. In addition to the direct requests, a department will receive requests from customers of type $i = 1, \ldots, n, i \neq j$, if an arc (i, j) exists in the network representing the system. All requests flowing in from arcs (i, j) such that $i \neq j$ are labeled as *alternate requests*, and will have rate $\lambda^a = \sum_{i \neq j} \lambda_i$. Thus, each department will receive direct requests as well as alternate requests, where the latter will be determined by the flexibility of the system. A revenue $\mathbf{r} = (r_1, \ldots, r_n)$ is associated with each customer request treated by department $i = 1, \ldots, n$. Each department decides which customer to accept or reject, given by the acceptance policy π. Then, the throughput of direct requests under a given acceptance policy will be denoted by $x(\pi)$ and the throughput for the total alternate requests by $y(\pi)$.

The network level analysis combines the lower level results of each department to obtain an upper bound on performance for the entire network. Define x_i as the flow (throughput) of direct requests through department i, and y_{ij} as the flow of alternate requests of type i that are routed through department j. Then, the network flow problem, that provides an upper bound on the performance of the entire system under any network level routing policy can be stated as

$$\max \sum_{i=1}^{n} [r_i(x_i + \sum_j y_{ij})]$$

$$\begin{aligned} \text{s.t.} \quad x_i + \sum_j y_{ij} &\leq \lambda_i \quad i = 1, \ldots, n & (1) \\ \sum_i y_{ij} &\leq M_j(x_j) \quad j = 1, \ldots, n & (2) \\ x_i \geq 0 \qquad y_{ij} &\geq 0 \quad i, j = 1, \ldots, n, & (3) \end{aligned}$$

where for each department j,

$$M_j(x_j) = M_j(\lambda_j^d, \lambda_j^a, C_j, x_j) = \max_\pi\{\mathbf{y}(\pi) : \mathbf{x}(\pi) > x_j\}$$

and represents the maximal mean departure rate (throughput) of alternate requests, subject to the requirement that the mean departure rate (throughput) of direct requests is at least x_j. For a proof of this result, the reader is referred to Theorem 2.1 in Kelly (1994).

Using this bounding scheme, and taking the value of the upper bound as a proxy for the throughput, we are able to provide the following structural result for a stochastic dynamic setting.

Theorem 4 *Consider a network $G(N, A)$ with symmetric demand vector $\boldsymbol{\lambda} = (\lambda, \ldots, \lambda)$, symmetric capacity vector $\mathbf{C} = (C, ..., C)$, symmetric revenue vector $\mathbf{r} = (r, \ldots, r)$ and routing vector $o_d(I)$. If $M(\lambda^x, \lambda^y, C, x)$ is increasing and concave as a function of λ^y, then whenever $G(N, A)$ does not have balanced skill diversity, one can find $G'(N', A')$ with identical demand vector $\boldsymbol{\lambda}$, capacity vector $\mathbf{C}$, routing vector $o'_d(I') = o_d(I)$, and $i'_d(J') \prec i_d(J)$ such that $T_{G'} \geq T_G$.*

It is known that (see Kelly, 1994 and references therein) this type of upper bound is tighter for well connected networks. We also know from earlier results that flexibility designs with chain like structures are superior to structures with no chains. Restricting the study of flexibility designs to such a domain ensures good connectivity in the underlying networks. Hence, we expect the proposed bounding scheme to perform well in the context of studying different flexibility designs.

Note that the fundamental structure of flexibility remains the same in the stochastic dynamic case. Given the nonlinear relationship between call volumes and capacity inherent in the network, we are unable to show the result for any $\boldsymbol{\lambda}$ and it is required that this vector be symmetrical. The numerical analysis will further explore this issue.

So far we have focused on flexibility design principles in relatively stylized settings. By moving from a deterministic capacity, discrete time period assumption to the stochastic dynamic setting in this section, our analysis has gained a lot of realism, however establishing general flexibility design principles has become more difficult. In order to test the applicability of our results for a real life business problem, we next explore the framework in a call center setting. The proposed implementation is subsequently used in the numerical analysis section. This illustration also allows us to demonstrate how the upper bounding technique can be used in a specific context.

4.2 An Application: Cross-Training in Call Centers

Let us consider a multi-department multi-skill call center whose flexibility structure is represented by the graph $G(N, A)$. In this context, if $(i, j) \in A$, service representatives of department j can handle requests of skill type-i. Let us denote by S_j the skill set of representatives of department j (i.e. $i \in S_j \iff (i, j) \in A$). Under this definition, a request of type i is a potential alternate request for department j $(i \neq j)$, if $i \in S_j$.

In order to assess the performance (expected throughput) in the stochastic-dynamic setting, we have to define precisely how dynamic routing will take place. The particular routing policy that will be employed is usually limited by technological constraints and may change from one call center to another. For the sake of concreteness, the following routing policy that is inspired from a real example is investigated. The center is designed such that, preferably, type-k requests are served at the department k even though they can be routed to any department j with the right competence $(k \in S_j)$. When a request of type-k arrives, it can be routed to any department j such that $k \in S_j$, and if there are representatives available at department j. If no customer representatives are available, the call is placed to the queue of department k (and has to be lost if the capacity of the queue is reached). As for service priorities, whenever a server of Department-k becomes available, service is immediately started on a request of type-k that is in the queue (of department k).

Assume that Department-k has C_k servers that all have identical skill sets. We denote by T_k the capacity of the queue of department k (excluding those customers already in service). Calls of Class-k arrive according to a Poisson process with rate λ_k. The service rate for skill-k requests at department-k is $\mu_{k,1}$ and the service rate for alternative requests at department k is $\mu_{k,2}$. The (exponential) abandonment rate for skill-k customers from the queue is θ_k.

The formulation of the exact optimal dynamic call routing is not presented since it will not be directly investigated. However, a brief discussion of some of the difficulties of a direct analysis is worthy. The state description of the above model requires three variables per department describing respectively the numbers of direct customers in service, alternate customers in service and direct customers in the queue. Even for a model with a few departments, the state space becomes extremely large. Given any particular configuration of flexibility, the optimal call routing policy can be obtained, in principle, as the solution of a stochastic dynamic program. The underlying controlled Markov process describing the number of requests in service of each type at each department is, however, multi-dimensional and an exact solution is impractical beyond single-department systems. Moreover, unless simulation is used, even computation of the performance measures under a suboptimal -but plausible- routing policy is difficult. The use of the bound for comparing different flexibility designs enables us to circumvent both difficulties. The approach yields an approximate performance measure under the best routing policy.

The bound of Section 4.1 utilizes some basic system parameters (such as arrival rates) as well as additional parameters that are computed from an underlying single department problem. The formulation of this single department problem and the details of the computation of the required parameters is presented in Appendix B.

5 Numerical Study

This section will explore the flexibility design principles from before, in a call center context. In particular, the bound of Section 4.1 is implemented for the model described in Section 4.2. Two sets of numerical examples are presented: one with a symmetric demand vector, and another with asymmetric demand. The systems have three departments and are symmetric in all the remaining parameters (i.e. $\mathbf{C}$, $\theta = (2, 2, 2)$, $\mathbf{T} = (25, 25, 25)$). All call types are assumed to have identical service rate $\mu = 1$ and identical revenues $r = 1$. For both sets of examples, we consider the eight system configurations depicted in Figure 5. Note that of these, configurations 1, 2, and 3 represent F_1, F_2, and F_3 respectively, while configurations 4-8 represent systems with some asymmetry either in their skill diversity or in their routing vectors.

Increasing and Diminishing Returns to More Flexibility: The results for the symmetric demand vectors are depicted in the graph in Figure 5. Throughput for each configuration is plotted for four different demand vectors, where the points have been joined for easier readability. The sum of the components of the skill diversity vector or the routing vector for each

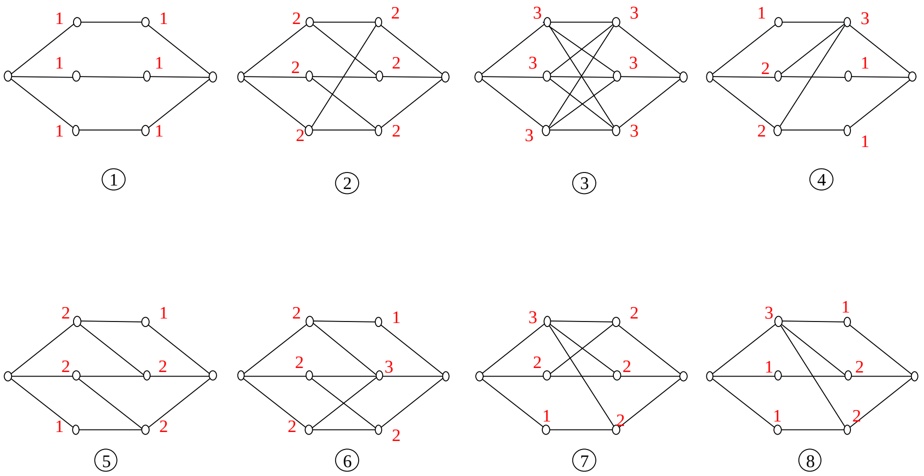

Figure 2: Eight flexibility configurations

configuration can be viewed as a measure for the total amount of flexibility in each system (3,6,9,5,5,6,6,5 for configurations one through eight). Using this quantification, note that as one would expect for systems with symmetric parameters, the examples with more flexibility exhibit higher throughput. Comparing the throughput for configurations 1,2, and 3, we note that throughput is increasing and concave in k for $k = 1, 2, 3$, as shown earlier for the case with deterministic capacity.

Higher Balance in Skill Diversity and Routing Implies Higher Throughput: Next, we explore the impact of different skill diversity and routing vectors on throughput performance. The first observation we make is that all the results are consistent with Theorems 3 and 4, as well as Corollary 2. Accordingly, configuration 5 with skill diversity vector $(1, 2, 2)$ is better than configuration 4 with skill diversity vector $(3, 1, 1)$. Similarly, configuration 2 with skill diversity vector $(2, 2, 2)$ is better than configuration 6 with skill diversity vector $(1, 3, 2)$. These cases illustrate examples where the skill diversity vectors are different and routing vectors are identical. Note that the routing vectors are not identical in the strict sense ((1,2,2) versus (2,2,1) for example), but given the assumptions for these examples, they are identical by symmetry. For the case where skill diversity vectors are identical and routing vectors differ, we can compare configurations 5 and 8, where once again configuration 5 with more balanced routing outperforms configuration 8. Similarly a comparison of configurations 2 and 7 is consistent with this result.

Comparing the Benefits of Routing Balance and Skill Diversity Balance: Comparing

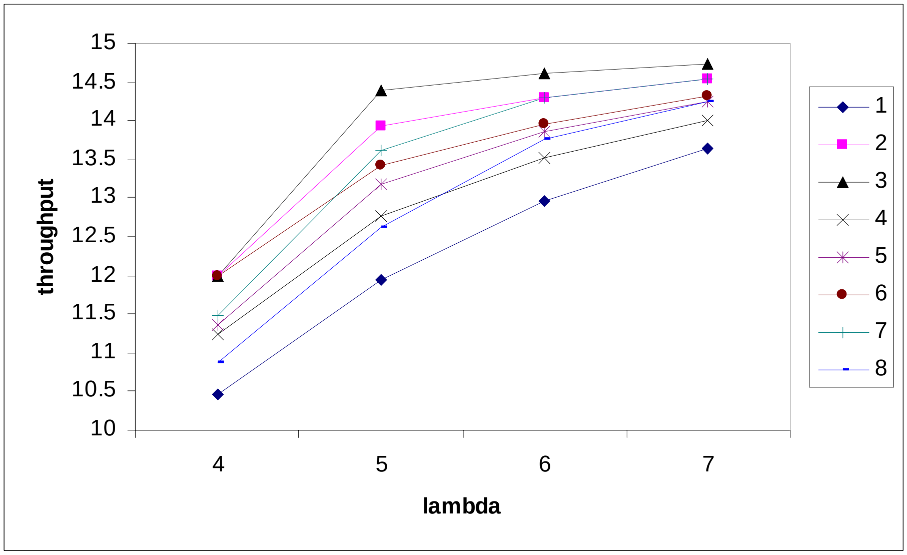

Figure 3: Throughput for symmetric demand vectors

configurations 6 with 7 and 4 with 8, we observe that both of these pairs represent cases where the skill diversity vector in one becomes the routing vector in the other and vice versa. Observe that configurations 6 and 4, with more balance in their routing vectors than in their skill diversity vectors are preferred for low values of λ, whereas configurations 7 and 8 with better balanced skill diversity and less balanced routing are preferred for higher λ. These comparisons illustrate an interesting point, namely that whether one prefers more balance in skill diversity or in routing depends on the load of the system. Intuitively, this reflects the fact that when capacity is the limited resource, skill diversity design is more important, whereas when demand is limited, balancing routing becomes more important.

The Interaction Between Flexibility and Capacity: In numerical examples (not reported here for brevity) where we vary capacity, first for fixed λ and then for λ such that the ratio $(\lambda/C\mu)$ remains constant, we make observations that are consistent with our expectations for these systems. For the former setting, we observe the same kind of qualitative change in the marginal value of capacity as a function of flexibility, as that stated in Theorem 2. The latter examples where utilization is kept constant demonstrate the well known scale effect. Thus we observe that *systems with smaller scale benefit more from flexibility design than systems that operate at a large scale.*

Asymmetric Demand: Figure 5 illustrates our results for asymmetric demand vectors.

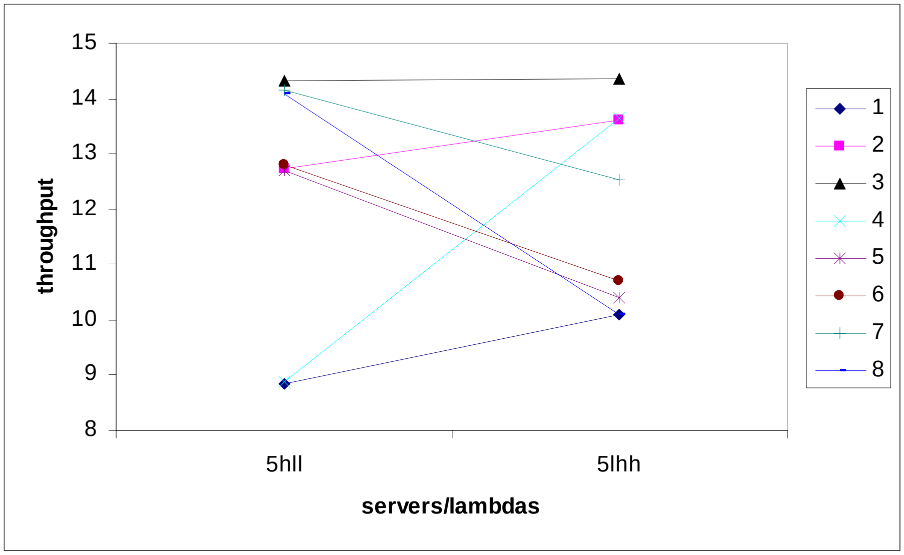

Figure 4: Throughput for asymmetric demand vectors

For these examples, the demand vectors are $HLL = (11, 2, 2)$, and $LHH = (1, 7, 7)$. Note that both of these demand vectors have a total demand flow rate of 15, but the demand is spread out differently in the two cases. A comparison of configuration 2 and 6, which have identical routing vectors, provides support for the balanced skill diversity result. An important difference to observe for these examples is that the asymmetry in λ has an impact on the ordering of the routing vectors (i.e. for example (1,2,2) and (2,1,1) are not equivalent anymore), and Corollary 2 which was shown for symmetric λ does not apply. Configurations 4,5,7, and 8 have different routing vectors, and exhibit different throughput performance with respect to each other depending on the demand vector. For example, configuration 4 which has fewer links than configuration 6, outperforms the latter for the demand vector LHH. This shows that *one can achieve better results with fewer links, for superior flexibility designs.* The example highlights the importance of the skill set design problem addressed in this paper. Our general results provide a basic understanding for good flexibility designs. The upper bound approach used herein can be used to make comparisons between specific configurations for more complex systems.

Summary of Results: The results from the numerical study can be summarized as follows. We are able to replicate most of the behavior shown for the system with deterministic capacity and discrete time periods, for the stochastic dynamic problem encountered in a call center. Our examples show that flexibility has diminishing returns to scale. The returns from flexibility interact with capacity, as characterized for the deterministic capacity case. The major difference between the deterministic and stochastic problems is the scale effect that is present in the latter. The numerical examples concur with the balanced skill diversity and routing guidelines shown for both deterministic and stochastic capacity systems. As call types become more diverse, it becomes difficult to generalize recommendations on flexibility design. However, our experience indicates that the approach in this paper provides an efficient and effective way to compare specific flexibility design alternatives. The reader is referred to Akşin and Karaesmen (2001) where we examine the usefulness of the approach in this paper as a comparison tool.

6 Concluding Remarks

In this paper, we study the problem of flexibility design for systems with multiple departments, multiple customer types, and service agents with multiple-skills. In this setting, we explore issues motivated by the questions of *how much flexibility* and *where and what type of flexibility* from a benefit standpoint. Our framework allows us to investigate these questions for increasing levels of realism, and provides both structural results and numerical support, in the presence of different assumptions pertaining to capacity and other system parameters. Our results show that the fundamental principles of flexibility are the same for systems with deterministic capacity and discrete time periods and those with queueing in continuous time. This suggests that it is the form of the network, rather than individual capacity or demand values that drive the qualitative behavior of flexibility value in such systems. This observation provides support for the approach in this paper, which approximates the lower level details and focuses on the network level effects of flexibility. The major difference between the systems with deterministic and stochastic capacity come from scale effects, which are not present in the case of a system with deterministic capacity. Systems with smaller scale will benefit more from a careful design of their flexibility levels and structures, compared to systems that operate at a larger scale.

We provide further evidence that smart limited flexibility is almost as good as full flexibility, and refine the description of what constitutes smart limited flexibility. A formal proof and characterization of the conditions under which a balanced skill diversity result or a balanced routing result (conjectured before in Jordan and Graves) holds is presented. By imposing a symmetric capacity condition (symmetric demand vector for the routing case), one is able to isolate the effect of different flexibility designs on performance. The result shows that in this kind of a setting, chain structures with more symmetry outperform those that have less symmetry, where this symmetry has been characterized by the skill diversity vector. Particularly for the case with deterministic capacity, this result is very strong and holds for any demand process. The numerical examples suggest that one needs to focus on skill diversity balance in high utilization systems, and on routing balance in low utilization systems.

What do our findings suggest for call center managers? The first implication is that for a given flexibility level, one prefers having spread out partial flexibility to focused flexibility with a select group of super-servers. This result holds for call center environments where the different departments are characterized by relatively similar parameters. For a multi-lingual call center, where the different departments in our model may represent different languages, this would be true, since one does not really expect huge asymmetries in call parameters for calls that on average only differ in the language dimension. One asymmetry that may be natural to expect in this kind of a multi-language setting is asymmetry in demand volumes, which we explored in the numerical analysis. We know from our structural results that for systems with deterministic capacity, the balanced skill set result holds irrespective of asymmetries in the demand vector. We expect this characteristic to be observed in large call centers. In general, our numerical examples highlight the point that routing vectors also have an impact on performance, and these have to be taken into account before making a definitive comparison between two systems that differ both in their skill diversity and routing vectors. Specific comparisons for more complex systems require further investigation either using the bounding approach of this paper or simulation. Based on observations from our numerical investigation, we can provide the following general guideline in this instance: a call center with heavier utilization should prioritize initiatives that lead to better balanced skill diversity, whereas those with lower utilization can focus on routing balance.

The question of whether one would rather have cross-trained servers throughout a service organization, or focus cross-training activities on an exclusive set of servers has been addressed from a human resource practice standpoint before. Hunter (1999) describes two prevailing models of work organization in retail banking (branch, call center), labeling one as the *inclusive* and the other as the *segmented* model. In terms of cross-training practice, the inclusive model implies cross-training for most employees throughout the organization whereas the segmented model refers to systems with cross-training for a select few. Based on an analysis of flexibility benefits only, our results provide evidence for the superiority of cross-training practices as found in the inclusive models of Hunter (1999), in systems with relatively low degrees of call diversity. As call diversity increases, this result may change in favor of cross-training choice that is more consistent with a segmented model. We further find that as call center scale increases, the importance of either skill set or routing design tends to diminish. Our scale related observations for flexibility design are consistent with the following empirical observation in Hunter (1999): *Smaller call centers were more likely to feature practices characteristic of the inclusive model. As traffic grows and call centers expand, segmentation appears to yield at least short term economic benefits.* The latter benefits come from cost factors not included in the analysis herein.

As suggested by our numerical examples with asymmetric demand volumes, the relationship between partial spread out flexibility and focused extensive flexibility is less clear for systems with complex structural or parameter related asymmetries. Differences in customer value or department capacities will further impact these types of comparisons. Earlier studies on optimal capacity investment in the presence of flexible resources (Van Mieghem 1998; Netessine et al. 2000) have shown that the value of flexibility depends on capacity choice and the two should be designed jointly. Linking the current analysis to the capacity optimization problem is the objective of ongoing research.

Acknowledgements: The authors would like to thank Lisa Fleischer, Itir Karaesmen, Ioana Popescu and Beril Toktay for helpful comments on the earlier versions of the paper.

References

[1] Akşin, O.Z. and Karaesmen, F. "Analyzing Flexibility in Call Center Design". *in preparation*, 2001.

[2] Bertsimas, D. and Chryssikou, T. "Bounds and policies for dynamic routing in loss networks". *Operations Research*, 47:3 379-394, 1999.

[3] Dawson, K. "CTI and the call center: the 2 % solution". *http://www.quicklink.com/ dawson/cticc.htm*, 1997.

[4] De Groote, X. "The flexibility of production processes: a general framework". *Management Science*, 40:7 933-945, 1994.

[5] Fine, C.H. and Freund, R.M. "Optimal investment in product-flexible manufacturing capacity". *Management Science*, 36 449-466, 1990.

[6] Gurumurthi, S. and Benjaafar S. “Modeling and Analysis of Flexible Queueing Systems”. *Working paper*, 2001.

[7] Harrison, J.M. and Lopez, M.J. “Heavy traffic resource pooling in parallel-server systems”. *Queueing Systems*, 33 339-368, 1999.

[8] Hopp, W.J., Tekin, E., and Van Oyen, M.P. “Benefits of skill chaining in production lines with cross-trained workers”. *Working paper*, 2001.

[9] Hunter, L.W. “Transforming retail banking: inclusion and segmentation in service work”. In P. Cappelli, ed., *Employment Practices and Business Strategy* Oxford University Press, New York, 153-192, 1999.

[10] Jordan, W.C. and Graves, S.C. “Principles on the benefits of manufacturing process flexibility”. *Management Science*, 41:4 577-594, 1995.

[11] Kelly, F.P. “Bounds on the performance of dynamic routing schemes for highly connected networks”. *Mathematics of Operations Research*, 19:1 1-20, 1994.

[12] Kelly, F.P. and Laws N.C. “Dynamic Routing in Open Queueing Models: Brownian models, cut constraints and resource pooling”. *Queueing Systems*, 13, 47-86, 1993.

[13] Koole, G. and Mandelbaum, A. “Queueing models of call centers: An introduction”. *Working paper*, 2001.

[14] Laws, C.N. “Resource pooling in queueing networks with dynamic routing”. *Advances in Applied Probability*, 24 699-726, 1992.

[15] Netessine, S., Dobson, G. and Shumsky, R. “Flexible service capacity: optimal investment and the impact of demand correlation”. *Operations Research, forthcoming*, 2000.

[16] Pinker, E. and Shumsky, R. “Efficiency-quality tradeoff of crosstrained workers”. *Manufacturing and Service Operations Management*, 2:1 , 2000.

[17] Sethi, A.K. and Sethi, S.P. “Flexibility in manufacturing: a survey”. *International Journal of Flexible Manufacturing Systems*, 2, 289-328, 1990.

[18] Sulkin, A. “ACDs: Heart of the customer contact center (still)”. *Business Communications Review*, 30:12, 46-50, 2000.

[19] Topkis, D.M. *Supermodularity and Complementarity.* Princeton University Press, Princeton, New Jersey, 1998.

[20] Turek, N. “Europe’s multilingual skills help it call around the world”. *Informationweek*, 809, 170, 2000.

[21] Van Oyen, M.P., Gel, E.G.S., and Hopp, W.J. “Performance opportunity for workforce agility in collaborative and noncollaborative work systems”. *IIE Transactions, forthcoming*, 1999.

[22] Van Mieghem, J.A. “Investment strategies for flexible resources”. *Management Science*, 44:1071-1078, 1998.

A Proofs

The following lemma is used in the proof of Theorem 1.

Lemma 1 *For the graph $G = (N, A)$ representing the flexible service system, any cut that satisfies one of the conditions below cannot be a minimum cut: i) There exists an $i \in X$ with $j \in A(i) \notin X$ ii) $f(X) \geq min(\sum_{i=1}^{n} \lambda_i, \sum_{i=1}^{n} C_i)$.*

Proof: Any cut X that satisfies i) will have $f(X) = \infty$. Since there exists cuts with finite capacity, X cannot be the minimum cut of this network. To show that any cut X that satisfies ii) cannot be a minimum cut, note that both $\sum_{i=1}^{n} \lambda_i$ and $\sum_{i=1}^{n} C_i$ are cuts of this network.

A.1 Proof of Theorem 1

The result will be shown for a network with $n = 3$. Let $\Delta_1 = T(F_2, \lambda, \mathbf{C}) - T(F_1, \lambda, \mathbf{C})$ and $\Delta_2 = T(F_3, \lambda, \mathbf{C}) - T(F_2, \lambda, \mathbf{C})$. Then the Theorem is equivalent to showing

$$0 \leq \Delta_2 \leq \Delta_1. \tag{4}$$

Recall that the maximum flow T_G for a network $G(N, A)$ is equal to the value of the minimum cut of G. This equality will be used in the ensuing proof.

First note that for the network with specialized servers, indexed by $k = 1$ and labeled as F_1, the minimum cut is given by

$$\sum_{i=1}^{n} \min(\lambda_i, C_i). \tag{5}$$

Similarly, for the fully flexible network with $k = n$ ($n = 3$) labeled as F_3, the minimum cut is given by

$$\min(\sum_{i=1}^{n} \lambda_i, \sum_{i=1}^{n} C_i). \tag{6}$$

In order to determine the values for Δ_1 and Δ_2, the minimum cut for F_2 needs to be established. The following eight cases are possible for a network with $n = 3$, where each case denotes the different possibilities for the componentwise minimum of λ and $\mathbf{C}$: Case 1:$(\lambda_1, \lambda_2, \lambda_3)$, Case 2:$(\lambda_1, \lambda_2, C)$, Case 3: $(\lambda_1, C, \lambda_3)$, Case 4:$(C, \lambda_2, \lambda_3)$, Case 5:$(\lambda_1, C, C)$, Case 6:$(C, \lambda_2, C)$, Case 7:$(C, C, \lambda_3)$, Case 8:$(C, C, C)$

For Case 1 and Case 8 $\Delta_2 = 0$ and $\Delta_1 \geq 0$ so the desired result in (4) holds. This condition needs to be verified for Cases 2 through 7 next.

Case 2: For this case $\lambda_1 \leq C$, $\lambda_2 \leq C$, and $\lambda_3 \geq C$. Using Equation 5, one then has that $T(F_1) = \lambda_1 + \lambda_2 + C$. For F_2, by Lemma 1, the five cuts that are candidates to be a minimum cut are $X^1 = \{s\}$, $X^2 = \{s, 1, 2, 3, 4, 5, 6\}$, $X^3 = \{s, 3, 4, 6\}$, $X^4 = \{s, 1, 4, 5\}$, $X^5 = \{s, 2, 5, 6\}$ and have values $f(X^1) = \lambda_1 + \lambda_2 + \lambda_3$, $f(X^2) = 3C$, $f(X^3) = \lambda_1 + \lambda_2 + 2C$, $f(X^4) = \lambda_2 + \lambda_3 + 2C$, $f(X^5) = \lambda_1 + \lambda_3 + 2C$.

Note that all other cuts are eliminated by one of the rules in Lemma 1. For Case 2, X^4 and X^5 cannot be the minimum cut since $\lambda_3 \geq C$. This leaves three cuts as candidates to be the minimum cut for F_2. If X^1 is the minimum cut, then $\Delta_1 = \lambda_3 - C > 0$ and $\Delta_2 = 0 < \Delta_1$, so the desired condition holds. If X^2 is the minimum cut, then $\Delta_1 = 2C - \lambda_1 - \lambda_2 > 0$ and $\Delta_2 = 0 < \Delta_1$, so the condition is again satisfied. For X^3 as the minimum cut, $\Delta_1 = C$ and $\Delta_2 = \lambda_3 - 2C$ if $T(F_3) = \lambda_1 + \lambda_2 + \lambda_3$, or $\Delta_2 = C - \lambda_1 - \lambda_2$ if $T(F_3) = 3C$. If X^3 is the minimum cut, then we further know that $f(X^3) < f(X^1)$ which implies that $\lambda_3 > 2C$. Similarly $f(X^3) < f(X^2)$ implies that $\lambda_1 + \lambda_2 < C$. These two conditions ensure that $\Delta_2 > 0$. If $\Delta_2 = \lambda_3 - 2C$, then jointly using the facts that $\lambda_3 \geq 2C$ and $\lambda_3 \leq 3C$, one can state that $\lambda_3 - 2C < C$, which implies that $\Delta_2 < \Delta_1$. If $\Delta_2 = C - \lambda_1 - \lambda_2$ then clearly $0 < \Delta_2 < \Delta_1$ which verifies the result for Case 2.

Case 3 and 4: The desired result holds by a symmetric argument to that in Case 2.

Case 5: For this case $\lambda_1 \leq C$, $\lambda_2 \geq C$, and $\lambda_3 \geq C$. Using Equation 5, one then has that $T(F_1) = \lambda_1 + 2C$. For F_2, by Lemma 1, the five cuts that are candidates to be a minimum cut are identical to those in Case 2 above, with the same values. This time, X^4 cannot be a minimum cut since both $f(X^3) \leq f(X^4)$ and $f(X^5) \leq f(X^4)$. This leaves four cuts as candidates to be the minimum cut for F_2. If X^1 is the minimum cut, then $\Delta_1 = \lambda_2 + \lambda_3 - 2C > 0$ and $\Delta_2 = 0 < \Delta_1$, so the desired condition holds. If X^2 is the minimum cut, then $\Delta_1 = C - \lambda_1 > 0$ and $\Delta_2 = 0 < \Delta_1$, so the condition is again satisfied. For X^3 as the minimum cut, $\Delta_1 = \lambda_2$ and $\Delta_2 = \lambda_3 - 2C$ if $T(F_3) = \lambda_1 + \lambda_2 + \lambda_3$, or $\Delta_2 = C - \lambda_1 - \lambda_2$ if $T(F_3) = 3C$. If X^3 is the minimum cut, then we further know that $f(X^3) < f(X^1)$ which implies that $\lambda_3 > 2C$. Similarly $f(X^3) < f(X^2)$ implies that $\lambda_1 + \lambda_2 < C$ and $f(X^3) < f(X^5)$ implies that $\lambda_2 < \lambda_3$. Hence, if $\Delta_2 = \lambda_3 - 2C$, then using the facts that $\lambda_3 \geq 2C$, $\lambda_3 \leq 3C$, and $\lambda_2 \geq C$ one can state that $\lambda_3 - 2C < C$, which implies that $\Delta_2 < \Delta_1$. If $\Delta_2 = C - \lambda_1 - \lambda_2$ then since $\lambda_1 + \lambda_2 \leq C$ and $\lambda_2 \geq C$, one has $0 < \Delta_2 < \Delta_1$ as desired. Finally, if X^5 is the minimum cut one has that $\lambda_3 < \lambda_2$, $\lambda_1 + \lambda_3 < C$ since $f(X^5) < f(X^2)$, and $2C < \lambda_2$ since $f(X^5) < f(X^1)$. This time, $\Delta_1 = \lambda_3$ and $\Delta_2 = \lambda_2 - 2C$ if $T(F_3) = \lambda_1 + \lambda_2 + \lambda_3$, or $\Delta_2 = C - \lambda_1 - \lambda_3$ if $T(F_3) = 3C$. Using the facts that $\lambda_3 < \lambda_2$, $2C \leq \lambda_2 \leq 3C$ and $\lambda_3 \geq C$, one can once again state that $\Delta_1 > \Delta_2$. In a similar fashion, if $\Delta_2 = C - \lambda_1 - \lambda_3$ then the fact that $\lambda_1 + \lambda_3 < C$ together with $\lambda_3 \geq C$ leads to the result that $\Delta_1 > \Delta_2$ which completes the verification of the result for Case 5.

Case 6 and 7: The result holds by a symmetric argument to that in Case 5.

A.2 Proof of Theorem 2

For the case when $\sum_{i=1}^{n} \lambda_i \geq \sum_{i=1}^{n} C_i$, first note the following equivalence. For any network G with demand vector λ, flexibility F_i and capacity vector $\mathbf{C}$, a corresponding network G' with demand vector $\lambda' > \lambda$, flexibility F_{i-1} and capacity vector $\mathbf{C}$ can be constructed such that $T_G(\lambda, C) = T_{G'}(\lambda', C)$ or equivalently $T(F_k, \lambda, C) = T(F_{k-1}, \lambda', C)$. More precisely, this construction can be performed by adding a vector $\mathbf{\Delta}^\lambda$ to the original demand vector λ, that ensures the equivalence in the maximum flows, with $\Delta_i^\lambda > 0$ only if i represents a department with excess capacity (i.e. $C_i > \lambda_i$). Making this transformation, the desired condition in the Theorem can be restated as $T(F_{k-1}, \lambda' + \Delta^C, C + \Delta^C) - T(F_{k-1}, \lambda', C) \geq T(F_{k-1}, \lambda, C + \Delta^C) - T(F_{k-1}, \lambda, C)$. Consider the following inequality, which is obtained by replacing $\lambda' + \Delta^C$ with λ'.

$$T(F_{k-1}, \lambda', C + \Delta^C) - T(F_{k-1}, \lambda', C) \geq T(F_{k-1}, \lambda, C + \Delta^C) - T(F_{k-1}, \lambda, C). \qquad (7)$$

Since T is non-decreasing in λ showing that this inequality holds is sufficient to show the desired result. It is known that the optimal value of the objective function in the minimum cut problem ($T(F_i, \lambda, \mathbf{C})$) is submodular in $(\lambda, -\mathbf{C})$ (Theorem 3.7.1 in Topkis, 1998). Combining this with the fact that for a function $f(x)$ that is submodular in x, $-f(x)$ is supermodular in x, one can state that $T(F_i, \lambda, \mathbf{C})$ is supermodular in $(-\lambda, \mathbf{C})$. Then the desired relationship in the inequality (7) holds by the definition of supermodularity, which completes the proof.

A.3 Proof of Theorem 3

Take any network, characterized by the graph $G(N, A)$. Let NA_G denote the set of cuts of this network, which cannot be eliminated by one of the rules in Lemma 1. This set will be called the set of uneliminated cuts. By Lemma 1, the minimum cut of the network G is a cut in NA_G. Now consider a second network $G'(N, A')$, where an arc (a, b) has been replaced by arc (a, b') and everything else is the same. The nodes $b \in J$ and $b' \in J'$ are chosen such that $i'_d(J') \prec i_d(J)$. Then according to Definition 3, the network G' is said to have more balanced skill diversity. It is next shown that G' thus obtained has maximum flow $T_{G'} \geq T_G$.

For any cut $X \in NA_G$ let X_I denote the nodes of X that are in the set I, i.e. $X_I = \{x \in X : x \in I\}$. Recall that $A(X_I)$ denotes the set of adjacent nodes to the nodes in X_I. By Lemma 1, all $y \in A(X_I)$ are also in X, i.e. $y \in X$. $NA_{G'}$ can be obtained from NA_G, by noting that the cuts in NA_G fall into three distinct sets: 1) The set of cuts that do not change as a result of the arc replacement being considered 2) The set of cuts that change, however have the same value as before 3) The set of cuts that are eliminated by Lemma 1 in the new network. In addition, there may be some new cuts.

More precisely, the cuts in NA_G can be grouped as follows. 1) All cuts $X \in NA_G$ such that $a \notin X$ belong to the first group, since the proposed change in the network only impacts $A(a)$. All such cuts will also be in $NA_{G'}$. 2) The cuts in the second group are those that are obtained from a cut $X \in NA_G$ by replacing the node b in the cut by b', i.e. $X' = X - b + b'$, such that for all $x \in X'_{I'}$ $A(X'_{I'}) \in X'$. Thus, these cuts cannot be eliminated by Lemma 1. Note that for this group of cuts, $f(X) = f(X')$ by symmetry of the capacity vector $\mathbf{C}$. 3) Consider a cut X with nodes $x_1 \in X_I$ and $x_2 \in X_I$, and $b \in \{A(x_1) \cap A(x_2)\}$. If for this cut $b' \not\in \{A(x_1) \cup A(x_2)\}$, then one will have $\{A'(x_1) \cup A'(x_2)\} \supset \{A(x_1) \cup A(x_2)\}$. Thus any cut $X \in NA_G$ with $x_1 \in X$ and $x_2 \in X$ will have $A'(X'_I) \supset A(X_I)$ in the new network G'. But then all of these cuts will be eliminated by condition i) of Lemma 1. 4) The argument for the third group of cuts shows that there may be some new cuts X' in $NA_{G'}$ with $x_1, x_2 \in X'$ and $A'(X'_I) \in X'$. In other words, these cuts X' contain all the nodes of cuts X that are in group three above, and in addition also contain node b'. Note that for these additional cuts $f(X) + C = f(X')$.

Characterizing the cuts of the new network, using the set of uneliminated cuts for the initial network, one observes that some cuts of G are eliminated in G', while those that are not eliminated preserve the same value $f(X) = f(X')$. All additional cuts that can be added to G' without being eliminated by Lemma 1 are shown to have $f(X') > f(X)$ for some $X \in NA_G, \notin NA_{G'}$. Thus one has that the minimum cut of G' is greater than or equal to the minimum cut of G, which implies that $T_{G'} \geq T_G$. The same argument can be repeated for another arc change that induces more balanced skill sets. Thus any network G' can be obtained from a network G through a finite number of arc changes, each improving throughput. This proves the result.

A.4 Proof of Corollary 2

Note that the service networks represented by graphs $G(N, A)$ are fully symmetric in λ and $\mathbf{C}$. In other words, a network where the flow is from the sink node towards the source node, and where the capacity vector $\mathbf{C}$ has been replaced by the demand vector λ and vice versa, has identical maximum flow with the original network. These latter types of networks can be labeled as the *reversed networks*. Observe, furthermore, that in the reversed network, all results previously shown for skill sets hold, and these are equivalent to results in terms of routings in the original network. Using this equivalence, and noting that the λ and $\mathbf{C}$ vectors in the corollary ensure that the reversed network has the same characteristics as the original network (any demand vector, symmetric capacity vector), the result stated in the Corollary follows by the proof for Theorem 3.

A.5 Proof of Theorem 4

The theorem is shown using the upper bound of the actual network, where revenues are set to one ($r = 1$). Take a network, characterized by the graph $G(N, A)$, such that the corresponding $M(\lambda^a, x)$ is concave increasing in the flow of total alternate requests λ^a. Now consider a second network $G'(N, A')$, where an arc (a, b) has been replaced by arc (a, b') and everything else is the same. The nodes $b \in J$ and $b' \in J'$ are chosen such that $i'_d(J') \prec i_d(J)$. Then according to Definition 3, the network G' is said to have more balanced skill diversity. It is next shown that G' thus obtained has throughput $T_{G'} \geq T_G$, where T_G is the throughput of the original network G.

An upper bound on T_G is given by the solution to the problem:

$$\max \sum_{i=1}^{n} [r(x_i + \sum_j y_{ij})]$$

$$\begin{aligned} \text{s.t.} \quad x_i + \sum_j y_{ij} &\leq \lambda \quad i = 1, \ldots, n && (8) \\ \sum_i y_{ij} &\leq M_j(\lambda_j^a, x_j) \quad j = 1, \ldots, n && (9) \\ x_i \geq 0 \qquad y_{ij} &\geq 0 \quad i, j = 1, \ldots, n. && (10) \end{aligned}$$

The upper bound on $T_{G'}$ is given by a similar problem, where the constraint (8) for $i = a$ is modified as

$$x_a + \sum_j y_{aj} - y_{ab} + y_{ab'} \leq \lambda \tag{11}$$

and the constraints (9) for $j = b, b'$ are modified as

$$\begin{aligned} \sum_i y_{ib} - y_{ab} &\leq M'_b(\lambda'^a_b, x_b) && (12) \\ \sum_i y_{ib'} + y_{ab'} &\leq M'_{\lambda'^a_{b'}, b'}(x_{b'}). && (13) \end{aligned}$$

Note that $M(\lambda^a, x)$ has been replaced by $M'(\lambda'^a, x)$ representing the different value this optimization problem will take once the arc change is performed in G, changing the alternate requests of department b and b' in the network. More precisely, for the symmetric case considered here, the proposed arc change will reduce λ_b^a by λ and increase $\lambda_{b'}^a$ by λ in the network G'. The symmetry of the revenue vector ensures that the two problems represented by G and G' have the same objective function. Furthermore, the symmetry of λ ensures that the constraint (8) has the same structure in both problems, despite the change in the constraint for $i = a$.

The desired result is shown by demonstrating that the upper bound problem for the network G' representing a problem with more balanced skill diversity, has a bigger feasible region (more relaxed set of constraints) and thus will have higher throughput. Let $x^\star = (x_1^\star, \ldots, x_n^\star)$ denote the optimum direct flows in the upper bound problem for G. Plugging the optimum value of x in constraints (8) then gives the constraints

$$\sum_j y_{ij} \leq \lambda - x_i^\star \quad i = 1, \ldots, n. \tag{14}$$

Let $xmin_{bb'}^\star = \min(x_b^\star, x_{b'}^\star)$. In the upper bound problem for G' impose the additional constraints that $x_b = x_{b'} = xmin_{bb'}^\star$. This implies that for the first set of constraints, the constraints for $i = b$ and b' become

$$\sum_j y_{ij} \leq \lambda - xmin^\star_{bb'} \quad i = b, b'. \tag{15}$$

Comparing (14) and (15) note that the corresponding y_{ij}s are less constrained in the latter. Next let us look at the second set of constraints in each problem. As stated earlier the arc exchange only impacts these constraints for $j = b$ and b'. For G these constraints will take the form

$$\sum_i y_{ij} \leq M_j(\lambda^a_j, x^\star_j) \quad j = b, b'. \tag{16}$$

Similarly, for the constrained version of G' one has

$$\sum_i y_{ij} \leq M'_j(\lambda'^a_j, xmin^\star_{bb'}) \quad j = b, b'. \tag{17}$$

Since the arc change in the original network is made such that the resulting skill diversity vector for G' is more balanced than that of G, we know by the symmetry of the demand vector that $\lambda^a_b > \lambda^a_{b'}$. The arc change implies $\lambda'^a_b = \lambda^a_b - \lambda$ and $\lambda'^a_{b'} = \lambda^a_{b'} + \lambda$. By the concavity of $M(x)$ in λ^a we then have that the increase $M'_{b'} - M_{b'}$ is bigger than the absolute value of the decrease $M'_b - M_b$. Combining this with the fact that $M(x)$ is also decreasing in x, it follows that by construction, the constraints (17) in the constrained version of the problem for G' are more relaxed than those in (16) for G. Since all the constraints in the constrained version of G' are either the same or more relaxed than those in G, this problem will have a higher throughput as the objective function value. Since this is true for a constrained version of G', it is also true for the original form of G', thus proving the desired result.

B The Single Department Problem

Let us now consider a single department -say department-k- which can handle two types of requests: direct and alternate. The department has a total of C servers with identical skill sets. Primary customers are placed in a queue of capacity T if there are no servers available. Both types of calls arrive according to Poisson process. The primary requests arrive at rate λ_1 and the secondary (alternative) requests at rate λ_2. The service times are assumed to be exponentially distributed with rate μ_1 for primary requests and rate μ_2 for alternate requests. As before, we denote by $x(\pi)$ and $y(\pi)$ the throughput of direct and alternate calls under a given routing policy π.

Recall now that the bound of Section 4 is based on the function $M(x)$ defined as

$$M(x) = \max_{\pi}\{y(\pi) : x(\pi) > x\}$$

Below, we express $M(x)$ as a linear program in which one of the constraints corresponds to $x(\pi) > x$. This representation leads to two important observations. First, $M(x)$ is decreasing (in a non-strict sense) and concave in x as x is the right-hand side of a convex optimization problem. Second, the constraint set $M(x)$ can be exactly characterized as the solution of a parametric linear program.

Let us denote by $q_1(t)$, $q_2(t)$ and $q_3(t)$ respectively the number of direct calls in service, the number of alternate calls in service and the number of direct calls in the queue. $A(t)$ denotes the control corresponding to the admission decision of the alternate calls. In particular, for any alternate request arrival instance t, $A(t) = 1$ corresponds to the decision to admit the alternate call and $A(t) = 0$ corresponds to the rejection decision.

The corresponding linear program is expressed in terms of the following decision variables: $q_{i,j,k} = P\{q_1(t) = i, q_2(t) = j, q_3(t) = k\}$ and $w_{i,j,k} = P\{q_1(t) = i, q_2(t) = j, q_3(t) = k, A(t) = 1\}$.

$$\begin{aligned}
\max\ y(\pi) \quad &= \quad \sum_{j=1}^{C} \mu_2 j \sum_{i=1}^{C} \sum_{k=1}^{T} q_{i,j,k} && (18)\\
\text{Flow In} \quad &= \quad \text{Flow Out} && (19)\\
\sum_{i=1}^{C} \mu_1 i \sum_{j=1}^{C} \sum_{k=1}^{T} q_{i,j,k} \quad &\geq \quad x && (20)\\
\sum_{i=1}^{C} \sum_{i=1}^{C} \sum_{k=1}^{T} q_{i,j,k} \quad &= \quad 1 && (21)\\
w_{i,j,k} \quad &\leq q_{i,j,k} && (22)\\
q_{i,j,k}, w_{i,j,k} \quad &\geq \quad 0 && (23)
\end{aligned}$$

Constraint set (19) corresponds to the flow balance constraints of the Markov Decision

Process and can be expressed as follows:

$$\begin{array}{ccc}
(\lambda_1 + i\mu_1 + j\mu_2)q_{i,j,k} + \lambda_2 w_{i,j,k} & = & \\
\lambda_1 q_{i-1,j,k} + \lambda_2 w_{i,j-1,k} + (i+1)\mu_1 q_{i+1,j,k} + (j+1)\mu_2 w_{i,j+1,k} & & \forall\, i,j,k : i+j < C \\
(\lambda_1 + i\mu_1 + j\mu_2)q_{i,j,k} & = & \\
\lambda_1 q_{i-1,j,k} + \lambda_2 w_{i,j-1,k} + (i+1)\mu_1 q_{i,j,k+1} + (j+1)\mu_2 w_{i-1,j+1,k+1} & & \forall\, i,j,k : i+j = C, k = 0 \\
(\lambda_1 + \lambda_2 + i\mu_1 + j\mu_2 + k\theta)q_{i,j,k} & = & \\
\lambda_1 q_{i,j,k-1} + i\mu_1 q_{i,j,k+1} + (j+1)\mu_2 q_{i-1,j+1,k+1} & & \forall\, i,j,k : i+j = C, k > 0, k < T \\
(i\mu_1 + j\mu_2 + k\theta)q_{i,j,k} & = & \\
\lambda_1 q_{i,j,k-1} & & \forall\, i,j,k : i+j = C, k = T
\end{array}$$

In order to ease the computational burden of solving the above parametric linear program an alternative approach can be taken. This consists of obtaining $M(x)$ by restricting the policy optimization to an easily parametrizable class of policies. In particular, a plausible class of policies is "server reservation" policies which are similar to 'trunk reservation' policies in communication networks. Within this class, the acceptance/rejection decision for alternative customers is determined by a single parameter n, in the following way. An alternative call is accepted if the total number of calls being serviced is less than or equal to n ($q_1 + q_2 \leq n$) and is rejected otherwise. In terms of the decision variables of the above LP, $w_{i,j,k} = 1$ when $q_1 + q_2 \leq n$ and $w_{i,j,k} = 0$ otherwise. The function $M(x)$ can then be obtained by varying the parameter n. As elaborated in Kelly (1994), this leads to an exact representation if the optimal policy for the single-department problem is indeed a server reservation policy. In our case, numerical results indicate that even though the server reservation policy is not exactly optimal for the single department problem, its performance is very close to optimal. Our numerical investigation is, therefore, based on server reservation policies and the function $M(x)$ is numerically computed for each parameter choice.