MODELING A PHONE CENTER: ANALYSIS OF A MULTI-CHANNEL MULTI-RESOURCE PROCESSOR SHARED LOSS SYSTEM

by

O. Z. AKSIN*
and
P. T. HARKER**

98/80/TM
(Revised Version of 97/57/TM)

* Assistant Professor of Operations Management at INSEAD, Boulevard de Constance, Fontainebleau 77305 Cedex, France.

** Department of Operations and Information Management, University of Pennsylvania, The Wharton School, Philadelphia, PA 19104-6366.

Printed at INSEAD, Fontainebleau, France.

Modeling a Phone Center:

Analysis of a Multi-Channel Multi-Resource Processor Shared Loss System

O. Zeynep Akşin * Patrick T. Harker †

December 1996, revised February 1998

Abstract

This paper presents a model for the study of operations at an inbound call center. The call center is modeled as a multi-class processor shared loss system, where the interacting effects of human, telecommunication and information technology resources are explicitly incorporated. Product form solutions for this type of system are provided along with expressions for performance measures like blocking and reneging. Some structural properties of system throughput are analyzed in an effort to pave the way for future optimization studies dealing with the design and management of phone centers.

*INSEAD, Technology Management Area, Boulevard de Constance, 77305 Fontainebleau Cedex, France; zeynep.aksin@insead.fr

†University of Pennsylvania, Department of Operations and Information Management, The Wharton School, Philadelphia, PA 19104-6366; harker@opim.wharton.upenn.edu

1 Introduction

Phone centers, also known as customer service centers and call centers, are becoming ever more prevalent in a variety of industries ranging from financial services, retail companies, to computer manufacturers. In fact, centralized call centers are hailed as one of the most cost-effective means of responding to customers Meehan (1993) . Gable (1993) asserts that for a majority of businesses, 70 - 95 % of customer contacts take place over the telephone. The recent decade has seen a growing focus on phone centers as alternative low cost service delivery channels across a variety of industries. This trend has been especially visible in the retail banking industry, where a move from traditional branch networks to service provision through remote service sites has been apparent. While this move has mostly been driven by a desire to lower service delivery costs, a new emphasis on sales and customer service along with intense competition with nonbank competitors has simultaneously increased the importance of phone centers as revenue generators. In designing and managing these centers, the best size and staffing levels have to be determined, along with the ideal balance between service and sales activities to ensure high profit generation. To achieve this type of performance, knowledge of the tradeoffs between cost cutting practices and revenue enhancement techniques is critical. The lack of a formal methodology that is capable of characterizing the economics in modern call center operations has instigated the ensuing research, wherein the nature of call center operations are studied in close detail.

The growth in the number of phone centers can, in part, be attributed to advances in telecommunication and information technology. In the past decade, automatic call dispatch-

ing systems have enabled firms to sort incoming calls so that they can be routed to the appropriate departments within a firm (Huffadine 1990; Hu 1990). With the current trend in computer telephone integration, these systems now have access to a firm's databases, LAN's (local area networks), and other management information systems, providing features like automatic number identification which pull-up a customer's files at a terminal before the agent picks up the phone. These and similar capabilities facilitate a variety of revenue generating transactions and sales which were traditionally not possible.

While there is a growing effort to fully automate some of the services provided at inbound call centers, a substantial number of transactions are still done through service agents interfacing between customers and technology. Quality of service is closely tied to the characteristics of the human and the information technology components of the resource base. A good example of the dependency between different types of resources is provided by the PRISM system at Merrill Lynch & Co., which allows a broker to answer questions immediately and enables almost any kind of transaction over the phone. For instance, a broker can enter an order, get it validated, transmit it to the appropriate exchange, have it executed and back at the terminal in less than a minute Kindel (1992).

In an environment where the variety of services and products offered are rapidly increasing, firms frequently provide specialized access to different types of customer demand by designating specialized service agents and phone lines to particular products. Thus, most of these centers have access through multiple channels, where each channel represents a group of similar products and services. While it is desirable from a service point of view to have spe-

cialized service agents, use of specialized information can frequently lead to inconsistencies within the system and difficulties in error resolution. At the same time, designing and maintaining individual information systems for each access channel is prohibitively expensive. To avoid these difficulties, firms have resorted to centralized information processing resources. Centralized information processing systems provide the desired consistency in data across functions within a call center and provide service representatives with efficient technological support. The systems enable access to information on a customer basis, and frequently allow the service representatives to make online changes to a customer's account. These changes are then automatically propagated through the system with the help of relational databases. As a result, while in a traditional call center service representatives would only be querying distributed databases, a phone center equipped with a customer information file-based system will have representatives constantly processing accounts on a central system. In the retail banking context, the consolidation of information systems within call centers as well as in branches, has led to the use of customer information files (CIF) or platform automation technology. The capacity implication of this new technology is a substantial load increase for the central processing equipment. In this type of a setting, it is essential to model the impact of the shared application on the performance of the system.

Thus, in a modern call center environment, capacity management translates into a complex process of managing the interaction between people and technology. It is the purpose of this paper to provide a performance model for call centers that enables a formal analysis of design and management questions emerging from a desire to become low cost, high

value added service delivery channels. The model captures some of the earlier mentioned characteristics of operations at these centers. In particular, it will incorporate the effect of consolidated information systems on service provision through multiple channels. A brief review of the related literature is provided in Section 2. This is followed by a formal statement of the model and its analysis in Section 3. Performance measures for the model are characterized in Section 4. Structural properties of system throughput are studied in Section 5. The paper concludes with a discussion of future research.

2 Related Literature

This section provides a brief review of the literature that is relevant to the ensuing modeling and analysis of the operations at a call center. The review starts out with background on the stochastic knapsack model and its analysis, which constitutes the basis of the results in Section 3. The section concludes with a sample of studies that report structural properties of performance measures in queueing and loss networks.

The stochastic knapsack model Ross and Tsang (1989), Ross and Yao (1990), Ross (1995) is a generalization of the Erlang loss system. In its basic form, this model consists of a given number of resource units, to which several classes of objects arrive. Each class of objects has its own arrival rate, mean holding time, and resource requirements. Interarrival times and holding times are distributed exponentially. When these assumptions are relaxed to allow for state dependent arrival and holding times, the system is known as a generalized stochastic knapsack. In both the basic and the generalized knapsacks, if upon arrival an

object finds fewer than the required number of resource units available, it is blocked and lost. The operations at a phone center are modeled as a generalized stochastic knapsack in what follows.

Recall that the information technology in a typical inbound call center is a resource that is jointly used by all servers of different specialization. To capture this characteristic in the model, a simple processor sharing service discipline for the phone center will be assumed (Kleinrock 1975). With this added feature, the model becomes a variant of the one described by De Waal and Van Dijk (1991), and De Waal (1993). Their model derives its motivation from a stored program controlled telephone switch application. While this paper will only consider a simple processor sharing scheme, the model by De Waal and Van Dijk allows for a class of different processor sharing schemes. The analysis focuses on a two class system whereas this paper intends to analyze the system for a general number of classes. Furthermore, the proposed model considers the additional variations where customers are put on hold and are allowed to renege from the system if their wait is too long. The only other published study to our knowledge that considers reneging behavior from a system point of view in a finite buffer system with processor sharing is the one by Coffman et al. (1994). Their model, however, does not account for multiple types of customers. A single class variation of the processor sharing system herein is also studied by Yamazaki and Sakasegawa (1987), where the impact of having a shared processor is analyzed. Foschini and Gopinath (1983) consider optimal acceptance policies to a three class processor shared system. Thus, the proposed model subsumes several characteristics that have appeared in

different works in the literature; namely, a system with access over multiple channels where calls are processed under a processor sharing discipline and the option of renegeing customers is considered.

It is known that both for these systems and for more general forms of a stochastic knapsack, calculation of the normalization constant that appears in the product-form solutions is not trivial (Kelly 1991). For real sized problems, blocking and reneging probability computations require special techniques that overcome this difficulty. For the model developed in this paper, methods that simplify the computation of performance measures are presented in Akşin and Harker (1997).

The natural extension of any performance evaluation endeavor is to use the description of existing performance in the system to improve performance through redesign. This type of an improvement is only possible if the designer is aware of some structural properties of performance measures that indicate the direction and the nature of the change in performance as a function of design parameters. Among the qualitative properties that have been identified as useful in designing these types of stochastic systems, one finds monotonicity of performance measures with respect to model parameters (see for example De Waal and Van Dijk 1991; Ross and Yao 1990) as well as second order properties like convexity or concavity (see for example Chang et al. 1991; Shantikumar and Yao 1989). Other structural properties of interest are stochastic orderings that provide bounds and inequalities, helping in the design of search heuristics for this class of problems (see for example Li 1994; Shantikumar and Yao 1988). The sequel will focus on establishing these types of structural properties for

total system throughput in inbound call centers.

3 The Performance Model

In this section, the operations of a phone center are modeled, providing a relationship between capacity choice and system performance measures, which can then be used to determine the relationship with system revenues. The model takes into account the uncertainty in demand, hence establishes a measure of capacity which explicitly deals with congestion. Capacity is a stochastic entity, which is a function of demand and resource allocation within the center. Resources that jointly determine capacity are human resources in the form of service agents, telecommunication resources as phone lines and VRUs (voice response units), and information technology resources. A customer call will require the availability of a phone line, through which the call can gain access to a service representative or a VRU. At the same time, the representative will need access to certain applications or databases in order to provide the requested services.

In the sequel, a phone center is modeled as a multi-channel queueing system with processor sharing. Each channel constitutes a department in the call center specializing on a specific set of products and services, and will alternately be called an access channel or department throughout the paper. Every channel will have a certain number of service representatives and phone lines associated with it. The specific assumptions underlying the proposed process model are described below. Figure 1 demonstrates the representation of a phone center. There are three different ways to measure performance, which are based on

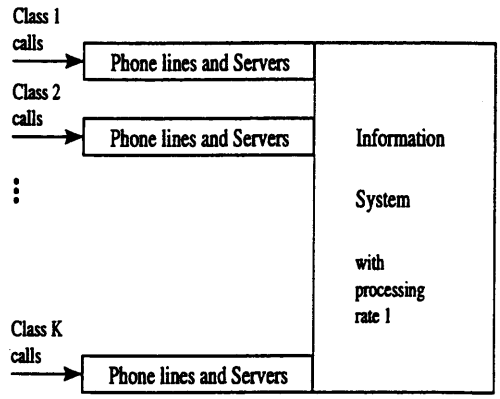

Figure 1: Representation of a Call Center

different assumptions regarding customer behavior and system configuration. In the most basic case, which is called the *loss system*, it is assumed that customers are extremely impatient. Hence, any customer who cannot initiate service immediately will leave. It is assumed that all customers who leave are lost demand and will not retry until their next transaction. In this configuration of the system, the number of trunks or phone lines are equal to the number of service representatives. Next, consider a system which may have phone lines in excess of the number of service representatives. Furthermore, drop the assumption of totally impatient customers. Upon arrival of a call, if all trunks are taken, the customer receives a busy signal and leaves the system. On the other hand, if a trunk is available but all agents are busy, the customer is put on hold and waits until an agent becomes available. This case will be called the *queueing system*. For many inbound call centers, the *system with reneges* will constitute the most realistic model. The system configuration is identical to the

queueing system described above. However, those customers that are put on hold do not necessarily wait until an agent becomes available. Some customers may exhibit impatience and leave the system while on hold before service initiation. This loss of customers is labeled as reneges.

Consider a phone center with K access channels. Each access channel consists of T_k, $k = 1, \ldots, K$ phone trunks and S_k, $k = 1, \ldots, K$ service agents specializing in product line k. For all three configurations of the system, one will have $T_k \geq S_k$. Customers arrive at the various access channels with an arrival rate of λ_k, where arrivals in each channel are independent of each other and the arrival process is assumed to be Poisson. Upon service initiation, the service representative will need access to the information system. This joint pool of information technology is capable of processing all transactions from different customers simultaneously. Notice that during times of high congestion, such central information systems respond with longer processing times. In other words, service times in the system are a function of the total number of customers being served in all channels. This characteristic is modeled as a processor sharing service discipline in what follows.

Let the information system be considered as a single server that processes at a constant rate of one service unit per unit time. Assume that each customer in class k with $k = 1, \ldots, K$ has a service requirement that is exponentially distributed with an average of $1/\mu_k$. That is, upon arrival to the system, each type of call will require a certain amount of processing. Since the speed of the processor is normalized at one, a faster processor will manifest itself as a smaller value for $1/\mu_k$ for the same type of call. The total processing time a call actually

experiences will depend on its service time requirement and the total number of customers present in the system at the same time. More specifically, letting $\mathbf{n} = (n_1, n_2, \ldots, n_K)$ denote the state vector, with n_k being the number of customers of class k in the system, the state dependent service rate for class k customers in the processor sharing loss system takes the form

$$\mu_k(\mathbf{n}) = \frac{n_k \, \mu_k}{(n_1 + \ldots + n_K)}. \tag{1}$$

One characteristic of this service rate is worth noting. The model assumes that a call will make continuous use of all three resources, i.e. the phone line, the service representative, and the information processing resource, throughout the duration of service. In other words, it is assumed that the information processing system processes a call for its entire duration. This assumption implies that the computer content of a call is equal to the labor content and the two happen simultaneously. While for the retail banking context this assumption would not distort reality very much, one can encounter other types of call centers where the computer content of a call is less than its labor content. If this were the case, then we would expect the processor sharing to have a lower impact on average talk times than what is implied by (1). Only that part of the call that involves the information processing resource would experience the impact of sharing. To illustrate how one can model this case, consider an instance where a class k call has computer content c_k, with $0 < c_k \leq 1$. If $c_k = 1$, state dependent service rates for class k will take the form in Equation (1). If this parameter is very close to zero, then the system will decouple into independent multiple server queues;

i.e. there will be no processor sharing. Otherwise, one has

$$\mu_k(\mathbf{n}) = \frac{n_k\ \mu_k}{1 + c_k(n_1 + \ldots + n_K - 1)}. \tag{2}$$

It will become evident in the sequel that this form for the state dependent service rate of class k calls renders the analysis of the model very difficult. In order to ensure tractability, it will be proposed to approximate (2) by

$$\mu_k(\mathbf{n}) = \frac{n_k\ \mu_k}{c_k(n_1 + \ldots + n_K)}. \tag{3}$$

For the remaining parts of this paper, the case with $c_k = 1$ will be analyzed. It will be assumed that the modification proposed in Equation (3) is used whenever this is not the case, merely by replacing μ_k by μ_k/c_k in the expression for $\mu_k(\mathbf{n})$. The sequel will illustrate that for large values of c_k, in other words when processor sharing has a significant impact, the system can exhibit some very unintuitive behaviour. One of the main contributions of the paper will be in illustrating this kind of qualitative behaviour and providing a methodology that enables the formal analysis of these types of systems.

To formalize the analysis of the proposed systems, one must introduce some additional notation. Let $X_k(t)$ denote the number of class k customers in the system at time t with $X(t) = (X_1(t), \ldots, X_K(t))$. Define $\pi(\mathbf{n})$ as the equilibrium probability of being in state $\mathbf{n}$ (i.e., of having n_k customers of class k in the system). Define the sets $\mathcal{A} = \{\mathbf{n} \in \mathcal{Z}_+^K : n_k \leq T_k\}$ and $\mathcal{A}_k = \{\mathbf{n} \in \mathcal{A} : n_k < T_k\}$, where $\mathcal{Z}_+$ denotes the nonnegative integers. Finally, $\mathbf{e}_k$ is a K-dimensional vector of zeros with a one in its kth position and $\mathbf{0}$ is a K-dimensional vector of zeros.

3.1 Determining Steady State Distributions

In order to characterize the performance of the phone center, one must establish the behavior of the system in steady state. To this end, one must first determine the equilibrium distributions, $\pi(\mathbf{n})$, for the three systems being considered. First, observe that all of these systems are generalized stochastic knapsacks with different state dependent arrival (for example $\lambda_k(\mathbf{n}) = \lambda_k 1(n_k < T_k)$) and service rates $\mu_k(\mathbf{n})$ (the reader is referred to Ross and Tsang 1989 and Ross 1995 for details on the stochastic knapsack problem). One can then use a result for the generalized stochastic knapsack, as shown in Ross (1995), to derive the equilibrium distribution. For completeness, this theorem is stated below. To derive equilibrium distributions, the theorem makes use of the concept of reversible stochastic processes. For a detailed exposition of this concept and related applications, the reader is referred to the book by Kelly (1979).

Theorem 1 *(Theorem 3.1 in Ross, 1995) For the generalized stochastic knapsack, a necessary and sufficient condition for* $\{\mathbf{X}(t)\}$ *to be reversible is that there exists a function* $\psi : A \rightarrow \Re_+$ *satisfying* $\psi(\mathbf{0}) > 0$ *and*

$$\frac{\lambda_k(\mathbf{n})}{\mu_k(\mathbf{n}+\mathbf{e}_k)} = \frac{\psi(\mathbf{n}+\mathbf{e}_k)}{\psi(\mathbf{n})} \qquad \forall\ \mathbf{n} \in \mathcal{A}_k,\ k = 1, \ldots, K. \tag{4}$$

Moreover, when such a function ψ *exists, the equilibrium distribution for the generalized knapsack is given by*

$$\pi(\mathbf{n}) = \frac{\psi(\mathbf{n})}{\sum_{\mathbf{n}\in\mathcal{A}} \psi(\mathbf{n})}, \qquad \mathbf{n} \in \mathcal{A}. \tag{5}$$

Note that the ratio $\frac{\psi(\mathbf{n}+\mathbf{e}_k)}{\psi(\mathbf{n})}$ can be interpreted as the likelihood of being in state $\mathbf{n}+\mathbf{e}_k$ with respect to the likelihood of being in state $\mathbf{n}$ in the long run.

3.1.1 The Loss System

The loss system is a generalized stochastic knapsack with state dependent arrival rates $\lambda_k(\mathbf{n}) = \lambda_k 1(n_k < T_k)$ and service rates $\mu_k(\mathbf{n}) = n_k\mu_k/(n_1 + \ldots + n_K)$, where $1(\cdot)$ is the indicator function. Recall from the initial description of the system that $S_k = T_k$, $k = 1, 2, \ldots, K$ for the loss configuration. Equation (4) takes the form

$$\frac{\psi(\mathbf{n}+\mathbf{e}_k)}{\psi(\mathbf{n})} = \frac{\lambda_k\,(n_1 + \ldots + n_K + 1)}{\mu_k\,(n_k + 1)},$$

which is clearly satisfied by

$$\psi(\mathbf{n}) = (n_1 + \ldots + n_K)! \prod_{k=1}^{K} (\frac{\lambda_k}{\mu_k})^{n_k} \quad \frac{1}{n_k!}. \tag{6}$$

Using equation (5), one obtains the equilibrium distribution as

$$\pi(\mathbf{n}) = \frac{1}{G}(n_1 + \ldots + n_K)! \prod_{k=1}^{K} \frac{(\rho_k)^{n_k}}{n_k!}$$

where $\rho_k = \lambda_k/\mu_k$ and $G = \sum_{\mathbf{n}\in\mathcal{A}} \psi(\mathbf{n})$, also known as the normalization constant.

3.1.2 The Queueing System

For this system, the definition of $\mathbf{n} = (n_1, n_2, \ldots, n_K)$ is modified such that it denotes the number of customers in the system including those that are waiting for a service representative to become available. Now $T_k > S_k$ for $k = 1, 2, \ldots, K$. Arrival rates still take the form

$\lambda_k(\mathbf{n}) = \lambda_k 1(n_k < T_k)$, and service rates are given by

$$\mu_k(\mathbf{n}) = \frac{\mu_k \ \min(n_k, S_k)}{\min(n_1, S_1) + \min(n_2, S_2) + \ldots + \min(n_K, S_K)}.$$

Equation (4) yields

$$\frac{\psi(\mathbf{n} + \mathbf{e}_k)}{\psi(\mathbf{n})} = \frac{\lambda_k(\min(n_1, S_1) + \ldots + \min(n_k + 1, S_k) + \ldots + \min(n_K, S_K))}{\mu_k \ \min(n_k + 1, S_k)}.$$

Using the convention that $\sum_a^b x = 0$ and $\prod_a^b x = 1$ if $b < a$, one obtains

$$\psi(\mathbf{n}) = (\sum_{k=1}^{K} \min(n_k, S_k))! \prod_{k=1}^{K} \left\{ \frac{(\rho_k)^{n_k} (\sum_{k=1}^{K} \min(n_k, S_k))^{(n_k - S_k)^+}}{(\min(n_k, S_k))! (S_k)^{(n_k - S_k)^+}} \right\}, \tag{7}$$

where

$$a^+ = \max(0, a).$$

The equilibrium distribution $\pi(\mathbf{n})$ can then be determined through equation (5).

3.1.3 The System with Reneges

The system with reneges is a slight modification of the above queueing system. The state vector $\mathbf{n}$ still denotes the number of customers in the system, including those that are waiting for an available agent. This time, one must model the customers that leave the system during their wait, before receiving any service. The literature on queueing systems provides two common approaches to modeling reneging behaviour of customers. Customers are assumed to renege with a certain probability, either based on the amount of time they spend waiting (see for example Abou-El-Ata and Hariri 1992; Montazer-Haghighi et al. 1986), or on the number of people in front of them in the queue (see for example Assaf and Haviv 1990;

Parkan 1987). Since the latter cannot be observed by customers of a phone center, the first approach will be adopted herein. In particular, the time a customer waits in queue k is assumed an exponential random variable with rate α_k. Customers are assumed to renege only when they are on hold, and will not renege once they start talking to a customer service representative. This implies a renege rate of $r_k(n_k) = \alpha_k(n_k - S_k)1(S_k < n_k \leq T_k)$ for $k = 1, \ldots, K$. Since reneging customers are the only difference between this system and the queueing system, state dependent arrival and service rates are the same. Define $\mu'_k(\mathbf{n})$ as the rate at which a customer of type k leaves the system when the system is in state $\mathbf{n}$. Since a customer only leaves as a result of a renege or at the time of service completion, this rate is clearly given by

$$\mu'_k(\mathbf{n}) = \mu_k(\mathbf{n}) + r_k(n_k).$$

To obtain an expression for the equilibrium distribution of this system, we replace $\mu_k(\mathbf{n} + \mathbf{e}_k)$ in equation (4) by $\mu'_k(\mathbf{n} + \mathbf{e}_k)$. Furthermore, for more compact notation, let $\tau_k(j, \mathbf{n}) = \mu_k \min(j, S_k) + r_k(j)(\sum_{k=1}^{K} \min(n_k, S_k))$. With some rearrangement, it can then be shown that

$$\psi(\mathbf{n}) = (\sum_{k=1}^{K} \min(n_k, S_k))! \prod_{k=1}^{K} \frac{\lambda_k^{n_k} (\sum_{k=1}^{K} \min(n_k, S_k))^{(n_k - S_k)^+}}{\prod_{j=1}^{n_k} \tau_k(j, \mathbf{n})}. \tag{8}$$

Using equation (5), the equilibrium distribution is given by

$$\pi(\mathbf{n}) = \frac{\psi(\mathbf{n})}{\sum_{\mathbf{n} \in \mathcal{A}} \psi(\mathbf{n})}.$$

4 Calculating Performance Measures

In order to determine revenue losses that result from congestion in the system, certain performance measures need to be established. Specifically, one would be interested in determining the probability of a customer being blocked upon arrival, as well as the loss that occurs due to reneging. In general, blocking probability in channel k is given by

$$B_k = 1 - \frac{\sum_{\mathbf{n}\in\mathcal{A}_k} \pi(\mathbf{n})}{\sum_{\mathbf{n}\in\mathcal{A}} \pi(\mathbf{n})}.$$

Note that obtaining these probabilities requires the calculation of a normalization constant G, which involves summing the expressions given in equations (6), (7), and (8), over a state space that can typically be very large.

Reneges are the second source of customer loss, so in addition to blocking probabilities, one needs to determine the portion of customers that are lost after entering the system. Denote the long-run probability of renege for a customer of type k by R_k. Then,

$$R_k = \sum_{\mathbf{n}\in\mathcal{A}} \pi(\mathbf{n}) \frac{r_k(n_k)}{\sum_{k=1}^{K}(\mu_k(\mathbf{n}) + r_k(n_k) + \lambda_k)}$$

which can equivalently be stated as

$$R_k = \frac{1}{G} \sum_{\mathbf{n}\in\mathcal{A}} \psi(\mathbf{n}) \frac{r_k(n_k)}{\sum_{k=1}^{K}(\mu_k(\mathbf{n}) + r_k(n_k) + \lambda_k)}. \tag{9}$$

The computation of R_k essentially involves a weighted sum of all the $\psi(\mathbf{n})$s, where the weights constitute the only difference between this and the computation of the normalization constant G.

Recognizing the difficulty in calculating these normalization constants, researchers have focused on the development of efficient methods for these calculations (see, for example,

Buzen 1973; Reiser and Kobayashi 1975; Lam and Lien 1983; Tsang and Ross 1990). In Akşin and Harker (1997), the authors develop methods to simplify the computation of performance measures for the model developed herein. It is clear that without these simplifying results, the model would not be useful in practice, since the calculation of performance measures for real-sized problems would constitute a problem that would not be computationally viable.

5 Qualitative Characteristics of System Throughput

In the previous sections, expressions for blocking and reneging probabilities were established. In the phone center context, blocked and abandoned calls translate into losses in revenue, since on average every customer call is a source of revenue for the firm. Hence, the calls that are not lost due to congestion related effects determine revenues for the center. More specifically, this implies that a knowledge of the characteristics of system throughput can be translated into a knowledge of the properties of average system revenues, assuming a given rate of revenue generation per customer. The earlier analysis indicates that throughput is a function of resource allocation, in particular of human resources, telecommunication resources, and information technology resources within the call center. Of these, the latter two are less flexible to reallocate with short notice, and new capacity additions in terms of these resources are typically not a matter of daily decision making. Human resources, on the other hand, are frequently used as a source of flexibility in call centers, resulting in various staffing practices. Though it is not the subject of the current paper to analyze these practices, the qualitative properties of throughput as a function of staffing will be explored

herein, to aid in the future study of these practices.

Propositions that establish structural properties of total revenues are stated below. The proofs for these are given in an Appendix that is available from the authors upon request.

5.1 Loss Case

Using notation from the above section, the throughput in channel k can be written as

$$TH_k = \lambda_k(1 - B_k(\mathbf{S}, \mathbf{T}, I)).$$

Now, assuming a fixed average revenue rate v_k for calls of type k, total revenues in the system can be expressed as a weighted sum of throughputs,

$$\sum_{k=1}^{K} v_k \lambda_k (1 - B_k(\mathbf{S}, \mathbf{T}, I)). \tag{10}$$

For simpler notation, let $w_i = v_i \lambda_i$ for $i = 1, \ldots, K$. Then,

Proposition 1 *(Server Allocation) Assume that the total number of servers $\sum_{k=1}^{K} S_k = S$ is fixed. For any class i and j with $w_i/\rho_i > w_j/\rho_j$ the optimal server allocation that maximizes the sum in (10) will have $S_i \geq S_j$.*

The proposition states an intuitive result; namely that those classes which bring in more revenue per work done should be assigned more servers. This type of an allocation tries to maximize total revenues generated in the system. A similar result has been shown to hold true for server allocation in a manufacturing system in Shantikumar and Yao (1988). The result is important, in that it significantly reduces the number of possible server allocations that need to be considered when staffing the center.

Definition 1 *(Chang et. al, 1991)* $\mathbf{X}(\theta)$ *is stochastically increasing (decreasing) and directional concave in sample path sense on* θ *if for every choice of* $\theta^{(i)}$, $i = 1,2,3,4$, *with*

$$\begin{aligned} \theta^{(i)} &\leq \theta^{(4)}, \qquad i = 1,2,3 \\ \theta^{(2)} + \theta^{(3)} &= \theta^{(1)} + \theta^{(4)}, \end{aligned}$$

there exists four random vectors $\hat{\mathbf{X}}^{(i)}$, $i = 1,2,3,4$, *defined on a common probability space, such that*

$$\begin{aligned} \hat{\mathbf{X}}^{(i)} &=_{st} \mathbf{X}(\theta^{(i)}), \qquad i = 1,2,3,4 \\ \hat{\mathbf{X}}^{(i)} &\leq \hat{\mathbf{X}}^{(4)} \ a.s. \qquad i = 1,2,3 \\ (\hat{\mathbf{X}}^{(i)} &\leq \hat{\mathbf{X}}^{(1)} \ a.s. \qquad i = 2,3,4) \\ \hat{\mathbf{X}}^{(1)} + \hat{\mathbf{X}}^{(4)} &\leq \hat{\mathbf{X}}^{(2)} + \hat{\mathbf{X}}^{(3)} \ a.s. \end{aligned}$$

Making use of this definition, we can state Proposition 2 as follows.

Proposition 2 *(Monotonicity and Concavity) The weighted sum of throughputs is stochastically increasing and directional concave in the sample path sense as a function of the server allocation vector* $\mathbf{S}$.

For a two class version of the loss model herein, De Waal and Van Dijk (1991) show that the throughput for a class of calls is monotonic in the number of servers allocated to that class. In other words, the throughput in Class 1 increases as one increases S_1 and holds S_2 constant, and decreases as one increases S_2 and holds S_1 constant. In Proposition 2, one is

interested in the monotonicity of the sum of throughputs in all classes. Given the result for the throughput of a particular class, it is not obvious whether the increasing effect in one class will dominate the decreasing effects in all other classes. Proposition 2 states that this is the case. This result will be useful in identifying server allocation schemes that maximize revenues in loss systems. In general, these kinds of monotonicity and concavity results are an essential prerequisite for any type of design endeavor.

5.2 The Queueing and Renege Cases

Numerical experimentation with the proposed process model suggests that the queueing system's behaviour is significantly more difficult to characterize than that of the loss system. This system seems to exhibit load dependent qualitative behaviour. Specifically, one observes that structural properties of total throughput vary as traffic intensity (λ/μ) or the number of calls that can be put on hold changes. Similar observations are made for the system with reneges, where the magnitude of renege rates are a third source of variation for the qualitative behavior of system throughput.

A numerical example is used to illustrate this point. The graphs in Figures 2 and 3 show total system throughput (Throughput) as a function of total number of servers (S) for the queueing system and the system with reneges respectively, in a hypothetical call center with three access channels. For each value of the total number of servers, only the throughput for the optimal allocation of the servers is plotted. Thus, for example, all different ways of allocating six servers to three classes are considered, and the one which results in the highest

Table 1: Parameters for queueing examples

Problem	$(\lambda_1, \lambda_2, \lambda_3)$	(μ_1, μ_2, μ_3)	(T_1, T_2, T_3)
a	(1,1,1)	(3,3,3)	(6,6,6)
b	(1,1,1)	(3,3,3)	(10,10,10)
c	(1,1,1)	(0.75,0.75,0.75)	(6,6,6)
d	(1,1,1)	(0.75,0.75,0.75)	(10,10,10)

total throughput is plotted as the point for six servers. This is done to ensure meaningful comparisons of total throughput, where the total number of servers are used in the best possible way within the center. These points have been joined with a continuous line for easier readability of the graphs. For simplicity, all examples depict a symmetric call center where each access channel is characterized by the same parameters.

The examples in Figure 2 compare the throughput in the queueing system for two different traffic intensity vectors and two different trunk size vectors. The parameters for the examples are tabulated in Table 1. The first thing to note is the jerky characteristic of the line that connects the points for different number of servers. It is clear from the graphs, that for these symmetrical systems, symmetrically allocated servers result in better total throughput compared to non-symmetrical allocations where one class has more servers allocated to it than another. For example we observe a dip in throughput in going from $S = 3$ to $S = 4$ in all four examples. This implies that the reduced blocking induced by the additional server in

one class is not enough to compensate for the increase in blocking experienced by the other two classes. This behaviour is clearly the result of processor sharing, where the number of servers allocated to a class determine the maximum number of customers of that class that can be served simultaneously, thus in a way determine the priority of a class with respect to processor utilization.

Example c illustrates another interesting property. For this example, in addition to the dips in throughput due to server allocation, the overall trend of total throughput as a function of total servers is also decreasing. This is a very surprising observation, given the conventional expectation that more servers will lead to higher throughput. Note, however, that the performance of the queueing system is determined jointly by server and information processing capacity. Example c depicts a case where the information processing resource is the bottleneck and adding more servers just increases the congestion experienced in processing. This in turn implies that there is a higher chance of having all trunks busy and hence an increase in blocking probabilities and a corresponding decrease in throughput. As shown in Example d, this problem of decreasing throughput can be overcome by adding phone lines to the system.

The examples for the queueing system are instructive in showing how the different resources interact to determine capacity. They also illustrate the importance of modeling the impact of processor sharing whenever it matters, since one observes qualitatively complex behavior that is hard to predict without a model. Next, the impact of customer reneges on this behavior is analyzed.

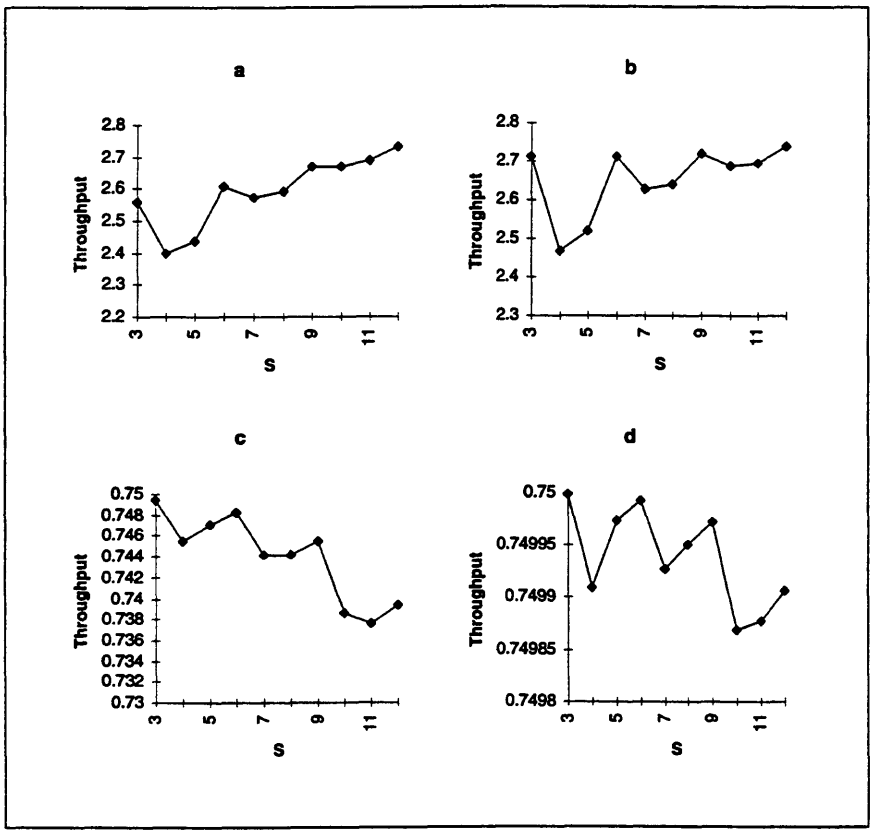

Figure 2: The Queueing System with (a) $\rho = (0.33, 0.33, 0.33)$ and $T = (6, 6, 6)$ (b) $\rho = (0.33, 0.33, 0.33)$ and $T = (10, 10, 10)$ (c) $\rho = (1.33, 1.33, 1.33)$ and $T = (6, 6, 6)$ (d) $\rho = (1.33, 1.33, 1.33)$ and $T = (10, 10, 10)$

Figure 3 depicts total throughput in the system with reneges. This time the vector of trunk sizes is fixed at $T = (6, 6, 6)$ for all four cases. However, two different renege rate vectors are considered. In Examples a and c renege rates are taken as $\alpha = (0.1, 0.1, 0.1)$ and in Examples b and d as $\alpha = (1.0, 1.0, 1.0)$. Examples a and b demonstrate the impact of renege rates on Example a for the queuing case. Similarly, Examples c and d illustrate the impact of renege rates on Example c in the queueing case. In addition to the earlier noted characteristics of total throughput as a function of total servers, one observes that a new tradeoff between losses from blocking and reneging is at work. Introducing reneging behavior reduces the blocking experienced in the system. This is obvious since the phone lines that were busy before could be available for a new arrival now if a customer lost patience and left the system. As one changes the total number of servers in this new system the loss from reneges will decrease since one will have less people on hold, however, this in turn will impact blocking probabilities as argued above. Thus, for the system with reneges it is clear that qualitative system behavior is further complicated by the tradeoffs between blocking and reneging. The examples clearly illustrate that changing the magnitude of the renege rate from 0.1 to 1.0 has an impact on the shape of the total throughput curve, supporting the point that structural properties for these systems are harder to establish and depend on the values the parameters take. The earlier conclusion, that interaction of different resources in determining capacity can lead to very unintuitive and hard to predict behavior in these systems is reinforced with these examples.

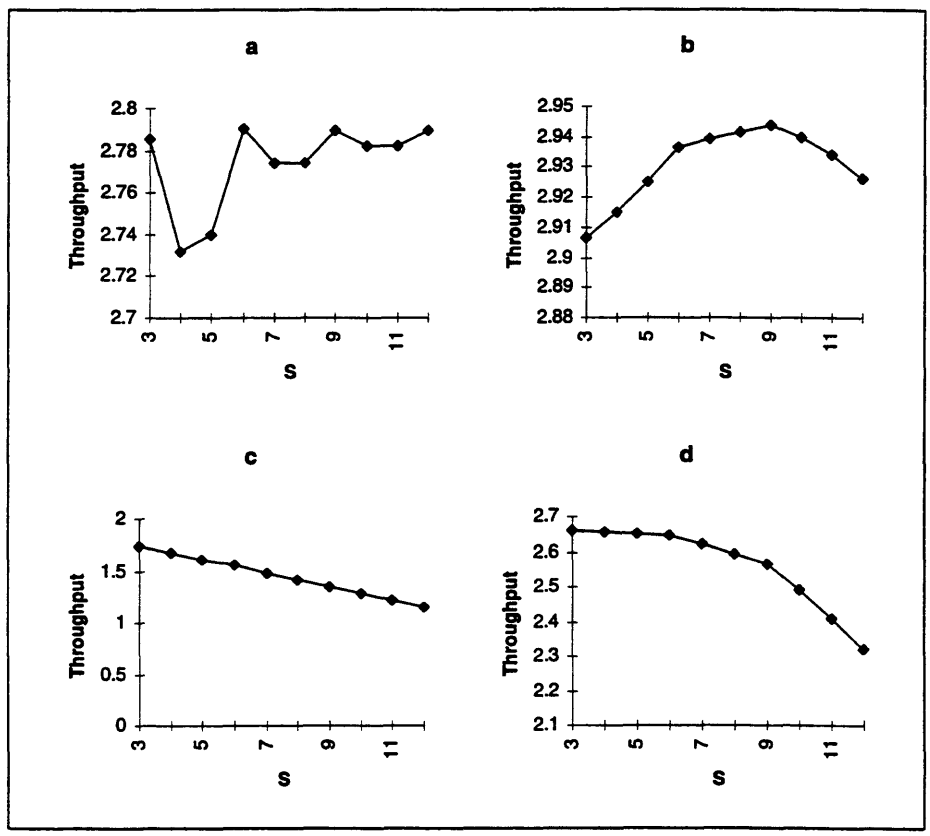

Figure 3: The Renege System with(a) $\rho = (0.33, 0.33, 0.33)$ and $\alpha = (0.1, 0.1, 0.1)$ (b) $\rho = (0.33, 0.33, 0.33)$ and $\alpha = (1, 1, 1)$ (c) $\rho = (1.33, 1.33, 1.33)$ and $\alpha = (0.1, 0.1, 0.1)$ (d) $\rho = (1.33, 1.33, 1.33)$ and $\alpha = (1, 1, 1)$

6 Concluding Remarks and Directions for Future Research

Motivated by issues arising from the design and management of modern call centers, this paper has presented a model and its analysis to characterize performance in these centers as a function of resource allocation. The model is the first in the literature to capture the impact of shared information processing resources on phone center performance. This characteristic is further enriched by explicitly modeling reneging behavior. The numerical examples in the preceding section illustrate how modeling these different characteristics can fundamentally change one's understanding of such systems, possibly leading to different managerial practices. While the model in this paper was motivated by phone centers in the financial services industry, one can think of other applications where it could be useful. Airline reservation systems constitute a good example for one such application where the central computing resource is essential in defining the capacity of a system.

The analysis of the model contributes to the literature in several dimensions. The product-form solutions derived for the queueing and reneging systems are the first ones reported for these types of systems. The results on the structural properties of system throughput in the loss system constitute an important contribution, both for the work related to call centers and for the design of data communication networks that share similar properties. The numerical examples illustrate the complex qualitative behavior that can be observed in such systems, and establish the exploration of structural properties for the

queueing and reneging systems as an important future research direction. In particular, these examples raise an interesting issue: When is it better to invest in additional service representatives and when would one be better off by investing in an information system upgrade? The model could be used to explore this important design tradeoff for call centers and similar systems.

The phone center performance model has emerged from an application in retail banking. In related research, the model is being used to analyze the transition to sales in traditionally service oriented call centers Akşin and Harker (1996). Use of the performance model enables a parametric analysis of different sales practices by characterizing service and sales tradeoffs, and establishing the costs and benefits involved. In Akşin and Harker (1996b), the performance model is embedded within an optimization problem that determines economically optimal staffing levels in a call center. The paper illustrates the potential use of the model in conjunction with design related issues by analyzing a very important design problem in call center management.[1]

References

[1] Abou-El-Ata, M.O. and Hariri, A.M.A. "The M/M/c/N queue with balking and reneging". *Computers and Operations Research*, 19:713–716, 1992.

[1]This research was funded by a generous grant from the Alfred P. Sloan Foundation. The authors would like to thank Keith W. Ross for his comments on earlier versions of this paper.

[2] Akşin, O.Z. and Harker, P.T. "To sell or not to sell: Determining the tradeoffs between sales and service in retail banking phone centers". Technical report, Financial Institutions Center, The Wharton School, 1996.

[3] Akşin, O.Z. and Harker, P.T. "Staffing a Call Center". Working Paper, 1996b.

[4] Akşin, O.Z. and Harker, P.T. "Computing performance measures in a multi-class, multi-resource, processor shared loss system". Working Paper, 1997.

[5] Assaf, D. and Haviv, M. "Reneging from processor sharing systems and random queues". *Mathematics of Operations Research*, 15:129–138, 1990.

[6] Buzacott, J.A., Shantikumar, J.G., and Yao, D.D. "Jackson network models of manufacturing systems". In D.D. Yao, editor, *Stochastic Modeling and Analysis of Manufacturing Systems*, pages 1–43. Springer-Verlag, New York, 1994.

[7] Buzen, J.P. "Computational algorithms for closed queuing networks with exponential servers". *Communications of the ACM*, 16:527–531, 1973.

[8] Chang, C., Chao, X., Pinedo, M. and Shantikumar, J.G. "Stochastic convexity for multidimensional processes and its applications". *IEEE Transactions on Automatic Control*, 36:1347–1355, 1991.

[9] Chang, C-S, Shantikumar, J.G., and Yao, D.D. "Stochastic Convexity and Stochastic Majorization". In D.D. Yao, editor, *Stochastic Modeling and Analysis of Manufacturing Systems*, pages 189–229. Springer-Verlag, New York, 1994.

[10] Coffman, Jr. E.G., Puhalskii, A.A., Reiman, M.I., Wright, P.E. "Processor-shared buffers with reneging". *Performance Evaluation*, 19:25–46, 1994.

[11] De Waal, P. "A constrained optimization problem for a processor sharing queue". *Naval Research Logistics*, 40:719–731, 1993.

[12] De Waal, P.R. and Van Dijk, N.M. "Monotonicity of performance measures in a processor sharing queue". *Performance Evaluation*, 12:5–16, 1991.

[13] Foschini, G.J. and Gopinath, B. "Sharing memory optimally". *IEEE Transactions on Communications*, COM-31:352–360, 1983.

[14] Gable, R.A. "Planning 800 disaster rerouting in distributed call centers". *Business Communications Review*, 23:74–77, 1993.

[15] Harris, C.M., Hoffman, K.L., and Saunders, P.B. "Modeling the IRS telephone taxpayer information system". *Operations Research*, 35:504–523, 1987.

[16] Hu, K.I. "Automatic call distribution system: A global tool". *CMA Magazine*, 64:8–11, 1990.

[17] Huffadine, R. "Trends in call management technology". *Telecommunications (International Edition)*, 24:57–58, 1990.

[18] Karlin S., and Proshan, F. "Polya type distributions of convolutions". *Annals of Mathematical Statistics*, 31:721–736, 1960.

[19] Kelly, F.P. *Reversibility and Stochastic Networks.* Wiley, Chicester, 1979.

[20] Kelly, F.P. "Blocking probabilities in large circuit-switched networks". *Advances in Applied Probability*, 18:473–505, 1986.

[21] Kelly, F.P. "Loss networks". *The Annals of Applied Probability*, 1:319–378, 1991.

[22] Kindel, S. "The telephone game ". *Financial World*, 161:74–75, 1992.

[23] Kleinrock, L. *Queueing Systems, Volume 2: Computer Applications.* Wiley, New York, 1975.

[24] Lam, S.S. and Lien, Y.L. "A tree convoluted algorithm for the solution of queueing networks". *Communications of the ACM*, 26:203–215, 1983.

[25] Li, H. "On a class of stochastic arrangement inequalities arising in optimal allocation of resources". *Probability in the Engineering and Informational Sciences*, 8:113–124, 1994.

[26] Meehan, T. "Voice processing improves services despite flaws". *Computing Canada*, 19:33–36, 1993.

[27] Montazer-Haghighi, A., Medhi, J. and Mohanty, S.G. "On a multiserver Markovian queueing system with balking and reneging". *Computers and Operations Research*, 13:421–425, 1986.

[28] Pacheco, A. "Second-order properties of the loss probability in M/M/s/s+c Systems". *Queueing Systems*, 15:289–308, 1994.

[29] Parkan, C. "Simulation of a fast-food operation where dissatisfied customers renege". *Journal of Operational Research Society*, 38:137–148, 1987.

[30] Reiser, M. and Kobayashi, H. "Queueing networks with multiple closed chains: theory and computational algorithms". *IBM Journal of Research and Development*, 19:283–294, 1975.

[31] Rohatgi, V.K. *An Introduction to Probability Theory and Mathematical Statistics.* John Wiley & Sons, New York, 1976.

[32] Ross, K.W. *Loss Models for Multiservice Telecommunication Networks.* Springer Verlag, 1995.

[33] Ross, K.W. and Tsang, D.H.K. "The stochastic knapsack". *IEEE Transactions on Communications*, 37:740–747, 1989.

[34] Ross, K.W., and Yao, D.D. "Monotonicity properties for the stochastic knapsack". *IEEE Transactions on Information Theory*, 36:1173–1179, 1990.

[35] Shaked, M. and Shantikumar, J.G. *Stochastic Orders and their Applications.* Academic Press, Inc., New York, 1994.

[36] Shantikumar, J.G. and Yao, D.D. "On server allocation in multiple center manufacturing systems". *Operations Research*, 36:333–342, 1988.

[37] Shantikumar, J.G. and Yao, D.D. "Second-order stochastic properties in queueing systems". *Proceedings of the IEEE*, 77:162–170, 1989.

[38] Tsang, D.H.K., and Ross, K.W. "Algorithms to determine exact blocking probabilities for multirate tree networks". *IEEE Transactions on Communications*, 38:1266–1271, 1990.

[39] Yamazaki, G. and Sakasegawa, H. "An optimal design problem for limited processor sharing systems". *Management Science*, 33:1010–1019, 1987.

Modeling a Phone Center:

Analysis of a Multi-Class Multi-Resource Processor Shared Loss System

APPENDIX

This appendix starts out with some definitions and preliminary results, and then proceeds with the proofs of Propositions 1 and 2.

Definition 2 *(Definition 1.4.1 in Buzacott et al. 1994) The* equilibrium rate *of X is a real valued, non-negative function, $r_X : \mathcal{N} \to \Re_+ = [0, \infty)$, defined as*

$$r_X(0) = 0, \qquad r_X(n) = p(n-1)/p(n), \quad n = 1, \ldots, N,$$

where $p(n) = P(X = n) > 0$ and X is a non-negative, integer valued random variable.

Based on the equilibrium rate, the PF_2 property (see Karlin and Proshan 1960) can be expressed as

$$X \in PF_2 \Longleftrightarrow r_X(n) \text{ increasing in } n. \tag{11}$$

Definition 3 *(Definition 1.4.6 in Buzacott et al. 1994) Let X and Y be two discrete random variables. Suppose their probability mass functions have a common support set $\mathcal{N}$. Let r_X and r_Y denote their equilibrium rates. Then $X \geq_{lr} Y$ if and only if $P(X = n)P(Y = n-1) \geq P(X = n-1)P(Y = n)$ for all $n \in N$, or equivalently, $r_X(n) \leq r_Y(n)$ for all $n \in N$.*

Lemma 1 *(Lemma 1.4.9 in Buzacott et al. 1994) Let $Y_1 \geq_{lr} Y_2$, and $Z \in PF_2$ is independent of Y_1 and Y_2. Then $Y_1 + Z \geq_{lr} Y_2 + Z$.*

Definition 4 *(Definition 1.1 in Li 1994)*

(1) Let $\mathbf{x} \in \Re^m$ and $\mathbf{y}$ be a permutation of $\mathbf{x}$. Then $\mathbf{x}$ is said to be more arranged than $\mathbf{y}$ (or $\mathbf{x}$ is less transposed than $\mathbf{y}$) if $\mathbf{x}$ can be obtained from $\mathbf{y}$ by a finite number of successive pairwise interchanges of two coordinates at a time such that each interchange results in an

increasing order for the interchanged elements. We denote this as $\mathbf{x} \geq_A \mathbf{y}$ *(e.g.,* $(4,5,3,1) \geq_A (4,3,5,1)$*).*

(2) A function $\phi : \Re^m \to \Re$ *is arrangement increasing (decreasing) if* $\mathbf{x} \geq_A \mathbf{y}$ *implies that* $\phi(\mathbf{x}) \geq (\leq)\phi(\mathbf{y})$.

Definition 5 *Let* $X = (X_i)_{i=1}^n$ *be a random vector, with its components not necessarily independent. Let* π *and* σ *denote two permutations of the* n *integers* $1, 2, \ldots, n$. *Let* X^π (X^σ) *denote a random vector obtained through rearranging the components of* X *according to the permutation* π (σ). *Define* $\pi \geq_A \sigma$ *following the same definition as for vectors. Then one can define a likelihood ratio ordering among the components of* X, *denoted* $X_n \geq_{lr:j} \ldots \geq_{lr:j} X_1$, *if* $E\phi(X^\pi) \geq E\phi(X^\sigma)$ *for all* $\pi \geq_A \sigma$ *and all arrangement increasing functions* ϕ.

Theorem 2 *(Theorem 5.4.4 in Chang et al. 1994) The components of* X *are ordered* $X_n \geq_{lr:j} \ldots \geq_{lr:j} X_1$ *if and only if the joint density function (or pmf) of* X *is an arrangement increasing function.*

Let $\mathbf{Y} = (Y_1, \ldots, Y_K)$, where $\{Y_k\}$ is a set of mutually independent random variables with

$$P(Y_k = n) = \frac{\rho_k^n}{n!} P(Y_k = 0), \qquad n = 0, 1, 2, \ldots.$$

Lemma 2 *If* $\rho_1 \leq \rho_2 \ldots \leq \rho_K$ *then* $Y_1 \leq_{lr} Y_2 \ldots \leq_{lr} Y_K$.

Proof By Definition 2, each Y_i has equilibrium rate

$$r_{Y_i}(n_i) = \frac{P(Y_i = n_i - 1)}{P(Y_i = n_i)},$$

which is equivalent to $r_{Y_i}(n_i) = \frac{n_i}{\rho_i}$. Then, for any index $i < j$ $\rho_i \leq \rho_j$, one has $r_{Y_i}(n) \geq r_{Y_j}(n)$. Using Definition 2, we get the desired result that $Y_i \leq_{lr} Y_j$. □

Lemma 3 *If* $Y_1 \leq_{lr} Y_2 \ldots \leq_{lr} Y_K$ *then* $X_1 \leq_{lr:j} X_2 \ldots \leq_{lr:j} X_K$.

Proof Make use of the product form solution

$$P(\mathbf{X} = \mathbf{n}) = \frac{(n_1 + n_2 + \ldots + n_K)! \prod_{k=1}^{K} (\frac{\rho_k^{n_k}}{n_k!})}{G}.$$

Take any index $i < j$. For fixed X_k, $k \neq i, j$, interchange the values of X_i and X_j in the pmf. Then, for any non-negative integer $x \geq y$, one obtains

$$\frac{P(\ldots, X_i = x, \ldots, X_j = y, \ldots)}{P(\ldots, X_i = y, \ldots, X_j = x, \ldots)} = \frac{P(Y_i = x)P(Y_j = y)}{P(Y_i = y)P(Y_j = x)} = \frac{r_{Y_j}}{r_{Y_i}} \leq 1$$

since $Y_i \leq_{lr} Y_j$. The above inequality implies that the joint pmf of X is an arrangement decreasing function. The desired result follows from Theorem 2. □

Lemma 4 $X_1 \leq_{lr} X_2 \ldots \leq_{lr} X_K$, *where the ordering* $\leq_{lr}$ *is concerned with the marginal distributions.*

Proof Using earlier notation, the equilibrium rate of the marginal distribution of X_1 can be derived as

$$r_{X_1}(n_1) = \frac{P(Y_1 = n_1 - 1)}{P(Y_1 = n_1)} \frac{\sum_{l=0}^{S-n_1+1} (l + n_1 - 1)![Y_2 \otimes \ldots \otimes Y_K](l)}{\sum_{l=0}^{S-n_1} (l + n_1)![Y_2 \otimes \ldots \otimes Y_K](l)},$$

which is clearly equivalent to

$$r_{X_1}(n_1) = r_{Y_1}(n_1)/r_{\Sigma_{(1)}}(S - n_1 + 1),$$

where $\sum_{(1)}$ denotes the sum of all Y_i, $i = 1, 2, \ldots, K$ excluding Y_1. To show $X_i \leq_{lr} X_j$, $i < j$, one needs to show that $r_{X_i} \geq r_{X_j}$, or that $r_{Y_i} \geq r_{Y_j}$ and $r_{\Sigma_{(i)}} \leq r_{\Sigma_{(j)}}$. $r_{Y_i} \geq r_{Y_j}$ is true by Lemma 2. From the definition of the likelihood ratio ordering, it is clear that showing $r_{\Sigma_{(i)}} \leq r_{\Sigma_{(j)}}$ is equivalent to showing $\sum_{(i)} = \sum_{(ij)} + Y_j \geq_{lr} \sum_{(ij)} + Y_i = \sum_{(j)}$. Since $Y_i \leq_{lr} Y_j$, by Lemma 1, all one needs to show is that $\sum_{(ij)}$ is PF_2. To show this, it is sufficient to note that $\sum_{(ij)}$ is a weighted convolution of Y_ks, where the Y_ks are PF_2 and the weights are log-concave in n. By Karlin and Proshan's result that the property of PF_2 is conserved under convolutions, the result follows. □

Recall the weighted sum of throughputs is given by,

$$\sum_{k=1}^{K} v_k TH_k(\mathbf{S}) = \sum_{k=1}^{K} v_k \lambda_k (1 - B_k(\mathbf{S}, I)), \tag{12}$$

and let $w_i = v_i \lambda_i$ for $i = 1, \ldots, K$.

Proof of Proposition 1 (Server Allocation) One proceeds by showing that for w_k/ρ_k ordered as indicated, and for S^π and S^σ as permutations of the allocation vector $\mathbf{S} = (S_1, S_2, \ldots, S_K)$, satisfying the condition $\pi \geq_A \sigma$, the following inequality holds:

$$\sum_k TH_k(\mathbf{S}^\pi) \geq \sum_k TH_k(\mathbf{S}^\sigma). \tag{13}$$

Using earlier notation, one can write the weighted sum of throughputs as

$$\sum_{k=1}^{K} (v_k TH_k) = \sum_{k=1}^{K} w_k \frac{G_k}{G}. \tag{14}$$

Initially, consider the case with $w_k = 1$ for all $k = 1, 2, \ldots, K$. Order the classes such that for $i < j$, $\rho_i \leq \rho_j$. Recalling by the definition of G_k that $\frac{G_k}{G}$ is the sum of the steady

state probabilities over all states where class k calls are not blocked, and making use of the probability axiom known as the *Principle of Inclusion-Exclusion* (Rohatgi 1976, page 27), the sum of the throughputs in equation (14) reduces with some algebra to

$$\sum_{k=1}^{K} TH_k(S_1, S_2, \ldots, S_K) = 2 - \sum_{k=1}^{K} P(X_k = S_k) + (-1)^K P(X_1 = S_1, X_2 = S_2, \ldots, X_K = S_K). \tag{15}$$

By Lemma 3, the joint probability distribution is an arrangement decreasing function. Also note that by the definition of marginal probabilities, $P(X_1 = S_1, X_2 = S_2, \ldots, X_K = S_K) \leq P(X_k = S_k)$ for all $k = 1, 2, \ldots, K$. Hence, to show that

$$\sum_{k=1}^{K} TH_k(S_1, S_2, \ldots, S_K) \geq \sum_{k=1}^{K} TH_k(S_2, S_1, \ldots, S_K) \tag{16}$$

for $S_1 \geq S_2$, it is sufficient to show that $P(X_1 = S_1) + P(X_2 = S_2) + \sum_{k=3}^{K} P(X_k = S_k) \leq P(X_1 = S_2) + P(X_2 = S_1) + \sum_{k=3}^{K} P(X_k = S_k)$. Equivalently, it is sufficient to show that

$$P(X_2 = S_1) - P(X_2 = S_2) \geq P(X_1 = S_1) - P(X_1 = S_2). \tag{17}$$

Multiplying the left hand side of equation (17) by $\frac{1}{P(X_2=S_1)}$ and the right hand side by $\frac{1}{P(X_1=S_1)}$, and noting that $\frac{1}{P(X_2=S_1)} \leq \frac{1}{P(X_1=S_1)}$ by Lemma 4 (hence preserving the inequality), the inequality in (17) reduces to

$$1 - r_{X_2} \geq 1 - r_{X_1}. \tag{18}$$

This is true by Lemma 4, so we have the desired result in (16). Since the arrangement ordering is defined through pairwise interchanges, it is sufficient to show the case for the interchange in (16), to verify (13) for the case when all $w_k = 1$. For the more general

situation when this is not the case, i.e. $w_k \neq 1$ for some k, we see that the above argument still holds, since the weights will cancel in (17) and all other results hold irrespective of the value of the weights. □

The following result is used in the ensuing proof of Proposition 2.

Lemma 5 *(Remark 2.7 in Chang et al. 1991) A function $f : (x_1, \ldots, x_n) \rightarrow \Re$ is directionally convex (concave) if and only if for every combination of four vectors $\mathbf{x}^{(i)} = (x_1^{(i)}, \ldots, x_n^{(i)})$, $i = 1, 2, 3, 4$, with*

$$\begin{aligned} \mathbf{x}^{(i)} &\leq \mathbf{x}^{(4)}, \qquad i = 1, 2, 3 \\ \mathbf{x}^{(2)} + \mathbf{x}^{(3)} &= \mathbf{x}^{(1)} + \mathbf{x}^{(4)} \end{aligned}$$

the inequality

$$f(\mathbf{x}^{(2)}) + f(\mathbf{x}^{(3)}) \leq (\geq) f(\mathbf{x}^{(1)}) + f(\mathbf{x}^{(4)})$$

holds.

The proof below shows the stochastically decreasing directional convexity of total blocking in the loss system, thus the stochastically increasing directional concavity of total system throughput, with respect to the server allocation vector $\mathbf{S}$. The result is shown in the sample path sense, but by Theorem 6.B.7 in Saked and Shantikumar (1994), it is known that this implies regular notions of stochastic convexity / concavity.

Proof of Proposition 2 (Monotonicity and Concavity) The proof uses standard coupling and uniformization techniques. The propositions are shown to hold for every choice of $\mathbf{S}^i$, $i = 1, 2, 3, 4$, satisfying

$$\begin{aligned} \mathbf{S}^{(i)} &\leq \mathbf{S}^{(4)}, \qquad i = 1, 2, 3 \\ \mathbf{S}^{(2)} + \mathbf{S}^{(3)} &= \mathbf{S}^{(1)} + \mathbf{S}^{(4)}. \end{aligned}$$

Consider four loss systems with K classes of customers, indexed by the superscript i, with server allocation vectors $\mathbf{S}^i$, $i = 1, 2, 3, 4$, as specified above. Let the service requirement μ_k of class $k = 1, \ldots, K$ customers be the same in all four loss systems. Let τ'_n, $n \geq 1$ be the arrival epoch of the nth customer in all four loss systems. Also let $\nu = \sup_{\mathbf{n},i} \sum_{k=1}^{K} \mu_k(\mathbf{n}, i)$, where $\mu_k(\mathbf{n}, i)$ denotes the state dependendent service rate for class k customers in network i when the state of the system is $\mathbf{n}$. Let $\{\tau''_n\}$ be the sequence of Poisson event epochs with rate ν, which are independent of $\{\tau'_n\}$ and let $\{\tau_n, n \geq 1\} = \{\tau'_n, n \geq 1\} \bigcup \{\tau''_n, n \geq 1\}$. Generate $\{U_n\}$, an iid sequence of uniform random variables on $[0, \nu]$, independent of $\{\tau_n\}$. Denote the number of customers in class j in the ith system at time τ_n as $X_j^i(\tau_n)$. Also let $B_j^i(\tau_n)$ be the cumulative number of blocked customers from class j, in system i, at time τ_n. Let $X_j^i(0) = 0$ and $B_j^i(0) = 0$ for all $j = 1, \ldots, K$ and $i = 1, 2, 3, 4$. To show that $\sum_{j=1}^{K} B_j^i(\tau)$ is decreasing and directionally convex as a function of the server allocation vector $\mathbf{S}$ in the sample path sense, one needs to show that the following conditions hold for all τ_n, $n \geq 1$:

$$\sum_{j=1}^{K} B_j^1(\tau_n) \geq \sum_{j=1}^{K} B_j^i(\tau_n) \quad i = 2, 3, 4 \quad n = 1, 2, \ldots \tag{19}$$

$$\sum_{j=1}^{K} B_j^1(\tau_n) + \sum_{j=1}^{K} B_j^4(\tau_n) \geq \sum_{j=1}^{K} B_j^2(\tau_n) + \sum_{j=1}^{K} B_j^3(\tau_n) \quad n = 1, 2, \ldots \tag{20}$$

In addition, the following condition on the states is imposed

$$X_j^1(\tau_n) \leq X_j^i(\tau_n) \quad j = 1, \ldots, K \quad i = 2, 3, 4 \tag{21}$$

Notice that all three conditions are satisfied at $n = 0$. Next assume that they hold for τ_n. To show that (19),(20) and (21) hold for all τ, one needs to show that the conditions still hold

at τ_{n+1}. The uniformized Markov chains $\sum_{j=1}^{K} B_j^i(\tau_n)$ and $X_j^i(\tau_n)$, $n \geq 1$, can be constructed as follows.

Case 1: Arrivals (in any class l, $l = 1, 2, \ldots, K$)

$$\sum_{j=1}^{K} B_j^i(\tau_{n+1}) = \sum_{j=1}^{K} B_j^i(\tau_n) + (X_l^i(\tau_n) + 1 - S_l^i)^+$$

$$X_l^i(\tau_{n+1}) = \min(X_l^i(\tau_n) + 1, S_l^i)$$

By construction, it is obvious that conditions (19) and (21) will be preserved at arrival epoch $n+1$. To show that

$$\sum_{j=1}^{K} B_j^1(\tau_{n+1}) + \sum_{j=1}^{K} B_j^4(\tau_{n+1}) \geq \sum_{j=1}^{K} B_j^2(\tau_{n+1}) + \sum_{j=1}^{K} B_j^3(\tau_{n+1}),$$

note this is equivalent to

$$\sum_{j=1}^{K} B_j^1(\tau_n) + (X_l^1(\tau_n) + 1 - S_l^1)^+ + \sum_{j=1}^{K} B_j^4(\tau_n) + (X_l^4(\tau_n) + 1 - S_l^4)^+ \geq$$
$$\sum_{j=1}^{K} B_j^2(\tau_n) + (X_l^2(\tau_n) + 1 - S_l^2)^+ + \sum_{j=1}^{K} B_j^3(\tau_n) + (X_l^3(\tau_n) + 1 - S_l^3)^+,$$

when the arriving customer is of class l. By the induction hypothesis, this reduces to

$$(X_l^1(\tau_n) + 1 - S_l^1)^+ + (X_l^4(\tau_n) + 1 - S_l^4)^+ \geq (X_l^2(\tau_n) + 1 - S_l^2)^+ + (X_l^3(\tau_n) + 1 - S_l^3)^+.$$

For notational convenience, let $\alpha_l^i = (X_l^i(\tau_n) + 1 - S_l^i)^+$. Consider the case

$$X_l^1(\tau_n) + X_l^4(\tau_n) \geq X_l^2(\tau_n) + X_l^3(\tau_n).$$

This implies

$$X_l^1(\tau_n) + X_l^4(\tau_n) = X_l^2(\tau_n) + X_l^3(\tau_n) + c,$$

where c is some non-negative constant. Now define $\bar{X}_l^3(\tau_n) = X_l^3(\tau_n) + c$, so that $X_l^1(\tau_n) + X_l^4(\tau_n) = X_l^2(\tau_n) + \bar{X}_l^3(\tau_n)$. This transformation preserves the condition in (21). Also let $\bar{\alpha}_l^3 = (\bar{X}_l^3(\tau_n) + 1 - S_l^3)^+$. Noticing that $\bar{\alpha}_l^3$ is directionally convex in $\bar{X}_l^3$ and $-S_l^3$ and making use of Lemma 5 and the induction hypothesis, it follows that

$$\alpha_l^1 + \alpha_l^4 \geq \alpha_l^2 + \bar{\alpha}_l^3,$$

or equivalently

$$\alpha_l^1 + \alpha_l^4 \geq \alpha_l^2 + \alpha_l^3 + c.$$

By construction

$$\begin{aligned}
\sum_{j=1}^{K} B_j^1(\tau_{n+1}) + \sum_{j=1}^{K} B_j^4(\tau_{n+1}) &= \sum_{j=1}^{K} B_j^2(\tau_n) + \sum_{j=1}^{K} B_j^3(\tau_n) + c_1 + \alpha_l^1 + \alpha_l^4 \\
&\geq \sum_{j=1}^{K} B_j^2(\tau_n) + \sum_{j=1}^{K} B_j^3(\tau_n) + \alpha_l^1 + \alpha_l^4 \\
&\geq \sum_{j=1}^{K} B_j^2(\tau_n) + \sum_{j=1}^{K} B_j^3(\tau_n) + \alpha_l^2 + \alpha_l^3 + c \\
&\geq \sum_{j=1}^{K} B_j^2(\tau_{n+1}) + \sum_{j=1}^{K} B_j^3(\tau_{n+1}),
\end{aligned}$$

where the first equality uses the induction hypothesis that

$$\sum_{j=1}^{K} B_j^1(\tau_n) + \sum_{j=1}^{K} B_j^4(\tau_n) = \sum_{j=1}^{K} B_j^2(\tau_n) + \sum_{j=1}^{K} B_j^3(\tau_n) + c_1,$$

with c_1 a non-negative constant. Now consider the case when

$$X_l^1(\tau_n) + X_l^4(\tau_n) \leq X_l^2(\tau_n) + X_l^3(\tau_n),$$

or equivalently

$$X_l^1(\tau_n) + X_l^4(\tau_n) + c = X_l^2(\tau_n) + X_l^3(\tau_n) \tag{22}$$

for a non-negative constant c. Proceed as before by defining $\bar{X}_l^4(\tau_n) = X_l^4(\tau_n) + c$. Notice that this transformation conserves the condition in (21). With the transformation, (22) becomes

$$X_j^1(\tau_n) + \bar{X}_j^4(\tau_n) = X_j^2(\tau_n) + X_j^3(\tau_n).$$

Again by the directional convexity of α_l^i and Lemma 5, one obtains

$$\alpha_l^1 + \bar{\alpha}_l^4 \geq \alpha_l^2 + \alpha_l^3, \tag{23}$$

or $\alpha_l^1 + \alpha_l^4 + c \geq \alpha_l^2 + \alpha_l^3$. Using a similar argument as before,

$$\begin{aligned} \sum_{j=1}^{K} B_j^1(\tau_{n+1}) + \sum_{j=1}^{K} B_j^4(\tau_{n+1}) &= \sum_{j=1}^{K} B_j^2(\tau_n) + \sum_{j=1}^{K} B_j^3(\tau_n) + \alpha_l^1 + \alpha_l^4 + c \\ &\geq \sum_{j=1}^{K} B_j^2(\tau_{n+1}) + \sum_{j=1}^{K} B_j^3(\tau_{n+1}), \end{aligned}$$

where the last inequality follows by (23). Thus (19), (20), and (21) are preserved at the arrival epochs. Now consider the case at departure epochs.

Case 2: Departures

$$\sum_{j=1}^{K} B_j^i(\tau_{n+1}) = \sum_{j=1}^{K} B_j^i(\tau_n)$$

$$X_l^i(\tau_{n+1}) = X_l^i(\tau_n) - \mathbf{1}_l^i \quad i = 1,2,3,4; \quad l = 1,\ldots,K$$

where

$$\mathbf{1}_l^i = 1\{U_n \in (0, \frac{\mu_l x_l^i}{\sum_{j=1}^{K} x_j^i}]\}.$$

The cumulative number of calls blocked in a class does not change at departure epochs. Hence, it is sufficient to show that (21) holds at departure epochs. This will insure that (19) and (20) hold for all $\tau_n, n \geq 1$ as shown in the case for arrivals.

In (21), consider the case when the inequalities are strict. Consider for example that $X_j^1(\tau_n) < X_j^2(\tau_n)$. By the definition of $\mathbf{1}_j^i$s, it is obvious that $\mathbf{1}_j^1$ is either 0 or 1 and similarly $\mathbf{1}_j^2$ is either 0 or 1. It is clear that any combination will insure $X_j^1(\tau_{n+1}) \leq X_j^2(\tau_n)$. Alternatively consider the case when $X_j^1(\tau_n) = X_j^2(\tau_n)$. By (21), it is also the case that $\sum_{j=1}^K X_j^1(\tau_n) \leq \sum_{j=1}^K X_j^i(\tau_n)$. These two conditions then imply that

$$\text{if } \mathbf{1}_j^1 = 1 \quad \rightarrow \quad \mathbf{1}_j^2 = 0 \text{ or } 1$$

$$\text{if } \mathbf{1}_j^1 = 0 \quad \rightarrow \quad \mathbf{1}_j^2 = 0.$$

Thus, $X_j^1(\tau_{n+1}) \leq X_j^2(\tau_{n+1})$. The cases for $X_j^1(\tau_{n+1}) \leq X_j^3(\tau_{n+1})$ and $X_j^1(\tau_{n+1}) \leq X_j^4(\tau_{n+1})$ are shown similarly. □