Control of Multiple Service, Multiple Resource Communication Networks

Scott Jordan, *Member, IEEE,* and Pravin P. Varaiya, *Fellow, IEEE*

Abstract— **The merging of telephone and computer networks is introducing multiple resources into networks, and information is becoming increasingly distributed across the network. Related services are being integrated onto a single network rather than being offered on separate uncoordinated networks. In this paper, we focus upon communication networks that integrate multiple services using multiple resources.**

In particular, we look at the decision of whether to accept or deny service requests in such a system. We prove a conjecture for the optimal policy for a related system introduced in [7] and characterize the optimal coordinate convex policy for our multiple service, multiple resource system.

I. Introduction

IN THIS PAPER, we focus upon communication networks that integrate multiple services using multiple resources. We investigate resource allocation strategies and try to capture the nature of controlling such a system. In particular, we prove a conjecture from [7] and then characterize the optimal coordinate convex policy for our multiple service, multiple resource model.

This work is motivated by several trends in networks. The merging of telephone and computer networks is introducing multiple resources into networks, and in formation is becoming increasingly distributed across the network. Related services are being integrated onto a single network rather than being offered on separate uncoordinated networks.

These trends are made possible by the availability of fiber and of inexpensive electronic storage, and by the introduction of greater intelligence into the signaling system. Furthermore, these trends are made profitable by the proliferation of desktop computers and the increased demand for better information transfer.

Proposals for implementing services in these *multiple service, multiple resource (MSMR)* networks abound. A few examples of these *services* might be electronic/voice mail, mixed media telephone calls, video conferencing, distributed databases, hypertext systems, electronic catalogues, electronic yellow pages, and collaborative editors.

Paper approved by I. Chlamtac, the Editor for Computer Networks of the IEEE Communications Society. Manuscript received April 16, 1991; revised July 27, 1992. This work was supported by the National Science Foundation under grants ECS-8719779 and ECS-8719298 and by the California State MICRO Program. S. Jordan was also supported by the Fannie and John Hertz Foundation. This paper was presented in part at IEEE INFOCOM, Bal Harbour, FL, April 1991.

S. Jordan is with the Department of Electrical Engineering, Northwestern University, Evanston, IL 60208 USA.

P. P. Varaiya is with the Department of Electrical Engineering, University of California at Berkeley, Berkeley, CA 94720 USA.

IEEE Log Number 9404727.

Our premise is that each *service* relies upon a number of underlying *resources* in the network. Examples of these *resources* might be communication links, databases, switches, storage devices, special purpose hardware and software. Although the precise meaning of "service" and "resource" and the relationship between them is a topic for future research, we assume in this paper that we have identified each service and the set of resources on which it depends.

Integrated services will share resources both for functionality and to decrease cost. Since these resources are limited, there will be interaction among the services. What types of interaction might we see? If you are the manager of a multiple service, multiple resource system, what requests for service do you accept? Based on what? If you base these decisons on maximizing revenue, what prices do you charge? And what resources should you acquire? The purpose of this research effort is to address such *resource allocation problems.*

In [8], we investigated the nature of this interaction. In this paper, we investigate the nature of controlling this type of system.

Considerable effort has been put into understanding related but simpler *multiple service, single resource* (MSSR) systems. In [1], Aein constructs a Markov chain model and states the resulting product form stationary distribution. Kaufman [10] shows that this product form holds under more general assumptions, including general service distributions. Foschini *et al.* [6] characterizes the optimal control policy among a wide class of policies for a two-service type one-resource type system. Ross and Tsang [24] extend this characterization for two-service type one-resource type systems to nonunit resource usage and to different arrival types. Ross and Yao [26] and Nain [20] investigate the effect of increasing traffic intensity upon throughput.

Some effort has also been applied to MSMR systems. In [14], [15], Kelly uses a MSMR framework to describe a circuit-switched network. He introduces the framework, states the stationary distribution, and obtains results relating to blocking probabilities, optimization and shadow prices by approximating the system as a collection of MSSR systems. In [4], Burman et al. obtain an insensitivity result for the stationary distribution of a MSMR system. Virtamo [28] displays a reciprocity relation in the sensitivity of blocking probabilities to traffic intensity. Numerical aspects are investigated in [5], [16], [19], [23], [25], [31], [33].

In addition, the MSMR system considered here is similar to some queueing systems. Foschini and Gopinath [7] investigate control policies to maximize throughput or minimize blocking

probabilities in a MSSR queueing system. Souza e Silva and Muntz [27] display sensitivity results for product form queueing systems.

In [8], we investigated the MSMR model for the simplest discipline: a request is granted if the necessary resources are available; otherwise it is rejected. In this paper, we consider strategies that may not accept all requests when the necessary resources are available, but may instead hold out for high paying requests. Section II displays the model originally presented in [8], and discusses classes of control policies. Section III introduces concepts relating to sets of states. In Section IV, we digress to a similar problem considered in [7] and prove a conjecture from that paper. Section V characterizes the optimal policy for our original problem and provides a few examples. In Section VI, we contrast this approach to the more general dynamic programming approach.

II. The Model

Consider the following model for resource allocation in *uncontrolled* multiple service, multiple resource (MSMR) communication networks.

Model: Consider a system that offers n types of services. Each service requires a set of resources (dependent upon the service type) to process. If these resources are available then the system manager accepts a service request, and then processing starts immediately; if the necessary resources are unavailable then the request is lost to the system. (In later sections, we will allow the system manager to deny a service request even if the necessary resources are available.)

Service requests arrive as independent Poisson processes. Each request occupies each resource that it needs for the same amount of time, and releases these resources simultaneously upon service completion. This amount of time is exponentially distributed, and independent of other service times.

We model this system as a Markov chain. Adopt the following notation.

A = a $m \times n$ matrix, with column i indicating the number of each of m resources used by service i.

b = a vector of length m indicating the number of each resource type in the system.

$\lambda \equiv (\lambda_1, \cdots, \lambda_n)$, the rates of incoming service requests.

$\mu \equiv (\mu_1, \cdots, \mu_n)$, the rates of service.

$\rho \equiv (\rho_1, \cdots, \rho_n)$, the loads, given by $\rho_i = \lambda_i/\mu_i$.

$L \equiv (L_1, \cdots, L_n)$, the rates of *accepted* service requests (throughput).

$x \equiv (x_1, \cdots, x_n)$, the state of the system where $x_i \equiv$ number of type i requests being processed.

$Z \equiv \{x \mid Ax \leq b$, i.e., x can be simultaneously processed with available resources$\}$.

$F_i \equiv \{x \mid x \in Z$ but $(x_1, \cdots, x_i + 1, \cdots, x_n) \notin Z\}$, the full set w.r.t. service type i.

$E_i \equiv \{x \mid x \in Z$ but $(x_1, \cdots, x_i - 1, \cdots, x_n) \notin Z\}$, the empty set w.r.t. service type i.

$\pi(x)$, the steady-state probabilities.

Our assumptions regarding the arrival and departure processes gives us a Markov chain on state space Z with transition rates:

$$r_{xy} = \begin{cases} \lambda_i, & \text{if } x \notin F_i \text{ and } y = (x_1, \cdots, x_i + 1, \cdots, x_n) \\ x_i\mu_i, & \text{if } x \notin E_i \text{ and } y = (x_1, \cdots, x_i - 1, \cdots, x_n) \\ 0, & \text{else.} \end{cases}$$

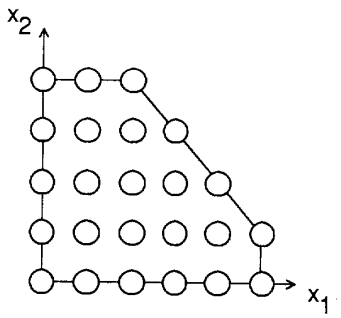


Fig. 1. The state space Z for a two service type, three resource type system.

Assume that service completion is never blocked. This implies that the state space Z is *coordinate convex*, i.e., if $x \in Z$ and $x_i \geq 1$, then $(x_1, \cdots, x_i - 1, \cdots, x_n) \in Z$.

The Markov chain is time reversible with well-known product form stationary distribution [4], [5]:

$$\pi(x) = \pi(0)\prod_{i=1}^{n}\frac{\rho_i^{x_i}}{x_i!} \qquad \pi(0) = \frac{1}{\sum_{x\in Z}\prod_{i=1}^{n}\frac{\rho_i^{x_i}}{x_i!}}. \quad (1)$$

As an example, consider a system that accepts only two types of service requests: type 1 requires one of resource A and one of resource B, and service type 2 requires one of resource B and one of resource C. If there are 5 A's in the system, 6 B's, and 4 C's, the state space Z would be as pictured in Fig. 1.

This model is discussed in more detail in [8].

A. Performance Measures

We now assume that the system manager can choose to deny a service request even if the corresponding resources are available. Why might she do this? Since resources are limited, accepting a request of one type may preclude the possibility of accepting a request of another type in the near future. Two measures are often used in resource allocation problems: throughput and blocking probability, see e.g., [6]. Fortunately, in our model, the two are equivalent for the purposes of control since they are linearly related by

$$L_i = \lambda_i[1 - \boldsymbol{P}(F_i)].$$

This suggests two measure of optimality:

1) $\min \sum_i c_i \boldsymbol{P}(F_i)$
2) $\max \sum_i r_i L_i$

where the $\{c_i\}$ and $\{r_i\}$ are costs and revenues correspondingly.

Although the two measures are equivalent through appropriate choice of $\{c_i\}$ and $\{r_i\}$, we find that they encourage different views of the system. Concentrating on *blocking probability* steers one to focus upon the effect of allowing or

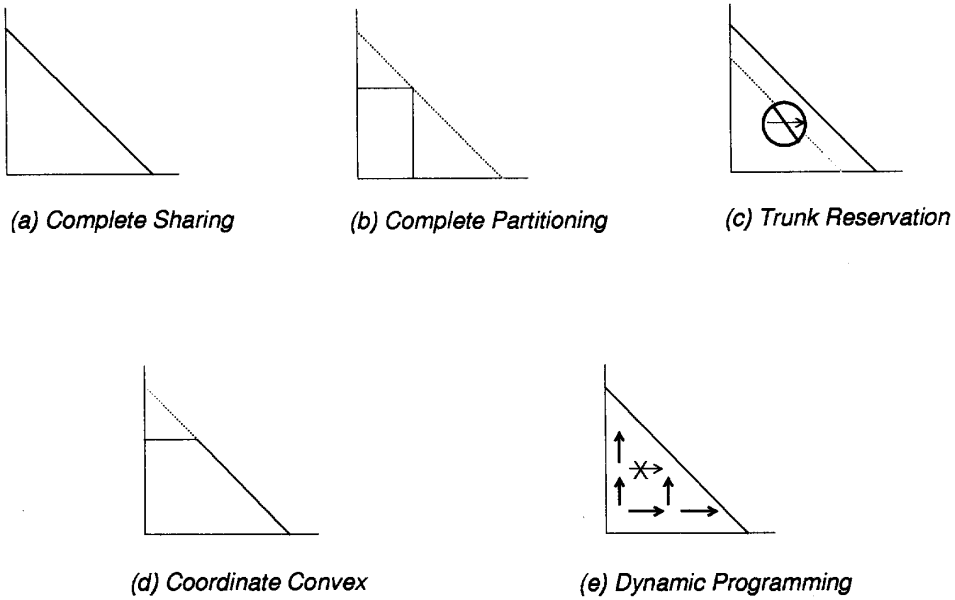


Fig. 2. Classes of control policies. (a) Complete sharing. (b) Complete partitioning. (c) Trunk reservation. (d) Coordinate convex. (e) Dynamic programming.

disallowing *transitions* in order to form desirable F_i sets. On the other hand, concentrating on *throughput* steers one to focus upon the value of *states*, or sets of states, toward the objective. This leads us to the next question: What do we know?

B. Information: Classes of Control Policies

On what information do you base your decision to accept or deny a service request? This depends both on what information you have available and on how much of that you choose to use. The first factor, availability, brings up monitoring and distributed system questions, all of which we are going to duck by assuming a full state information centralized control mechanism. The second factor affects the complexity of the control. The more information we use, the tougher the control policy is to implement. But closer to a personal point, the more information we use, the tougher it is to characterize any form of an optimal policy.

To help illustrate some possibilities, consider the simplest MSMR system with sharing: one with two services and one resource type. This system, if uncontrolled, has a triangular state space as pictured in Fig. 2(a).

The simplest type of control, namely no control at all, is to share the resource completely between the two services [Fig. 2(a)]. This requires no information. The opposite strategy is to completely partition the resources between the two services [Fig. 2(b)]. This is equivalent to restricting the state to some rectangular subset of the original triangular space, and requires only separate knowledge of the number of each service currently in the system.

A popular control class in telephone networks is trunk reservation [Fig. 2(c)]. In this type of policy, all arrivals of one type are accepted, but arrivals of the other type are accepted only if there are at least some minimum number of idle resources. This requires combined state information, but bases control on only one parameter, the minimum number of idle resources.

A more general class of policies than complete sharing or complete partitioning is coordinate convex (c.c.) policies [Fig. 2(d)]. In this class, admission decisions depend on *the state the system would enter if the request is granted.* This is equivalent to *restricting the state to some subset of the original space.* Since service completion can not be realistically blocked, this subset, like the original state space, must be coordinate convex. This class includes the complete sharing and complete partitioning classes as subsets and also allows for policies that reserve some resources for each service type and share the rest.

The most general class of policies base decisions *not only upon what state admission would place the system in, but also upon what type of service is requested* [Fig. 2(e)]. This corresponds to allowing or disallowing each upward transition in the Markov chain, and includes all of the above classes as special cases. The only approach we are aware of that lends insight into when to disallow individual transitions, however, is dynamic programming, see, e.g., [23]. Although this is a useful numerical technique, little ground has been made toward characterizing optimal policies.

We therefore back off to coordinate convex policies. This class has the nice property that the model of the controlled system is the restriction of the uncontrolled time reversible Markov chain to a subset, and is thus itself time reversible, maintaining its nice product form stationary distribution. Coordinate convex policies have thus been a popular class of control policies, see e.g., [6], [24].

Having chosen this class of policies, we return to the question of the previous section: what performance measure do

we use? Since *c.c. policies can be represented by a set of states*, we choose the approach that lends itself to focusing upon states rather than transitions. The following notation helps:

$r_i \equiv$ revenue generated by servicing request type i, per unit of time.

$r(x) \equiv$ rate of revenue generated while in state x, namely $\sum_i r_i x_i$

$R \equiv Er(X) = r(EX)$

$=$ the average revenue per time unit generated by the system.

Our objective is then to choose the c.c. subset $Z^* \subseteq Z$, that maximizes $E[r(X) \mid X \in Z^*]$.

III. Set of States

To talk about sets of states it is helpful to introduce some constructs that relate sets to each other:

$E_V r(x) \equiv E[r(X) \mid X \in V]$ where $V \subseteq Z$

V is annexable to Z iff $V \cap Z = \varnothing \& V \cup Z$ is c.c.

V is removable from Z iff $V \subseteq Z \& Z - V$ is c.c.

Annexable and removable sets were introduced by Foschini in [6], [7], initially as "incrementally admissible" and "incrementally removable" sets.

Since the Markov chain is time-reversible, the removal of a set V from Z affects the distribution on $Z - V$ only through the normalization constant $\pi(0)$. Therefore, $E_V r(x)$ does not depend on the policy Z^*. Furthermore, the removal of V from Z increases the average revenue on Z only if $E_V r(x) < R$. We can therefore characterize an optimal c.c. policy as a subset of Z to which nothing above average can be added and nothing below average removed [6], i.e.,

A c.c. set $Z^* \subseteq Z$ is optimal iff:

$\nexists V \subseteq Z \ni V$ is annexable to Z^* & $E_V r(x) > R^*$

$\nexists V \ni V$ is removable from Z^* & $E_V r(x) < R^*$

where $R^* \equiv E_{Z^*} r(X)$.

Partition Z into those states generating an above average rate of revenue, $Z^+ \equiv \{x \mid x \in Z \ \& \ r(x) > R\}$ and those generating a below average rate of revenue, $Z^- \equiv \{x \mid x \in Z \ \& \ r(x) \leq R\}$. Our control problem is to keep states in Z^+ and get rid of states in Z^- such the remaining set is c.c. and generates the highest possible average rate of revenue.

Define the supporting set of V to be the minimum set required to make V c.c., namely,

$$\mathrm{ss}(V) \equiv \{x \notin V \mid 0 \leq x_i \leq v_i \quad \forall i \text{ for some } v \in V\}.$$

Consider a c.c. set $\hat{Z} \subset Z$.

Certainly we should remove from $\hat{Z}$ any states in $\hat{Z}^-$ that do not support states in $\hat{Z}^+$, namely,

$$\mathrm{junk}(\hat{Z}) \equiv \hat{Z}^- - \mathrm{ss}(\hat{Z}^+)$$

since $\mathrm{junk}(\hat{Z})$ is removable from $\hat{Z}$ and since $\mathrm{junk}(\hat{Z})$ generates a below average rate of revenue, namely, $E_{\mathrm{junk}(\hat{Z})} r(X) < \hat{R}$ where $\hat{R} = E_{\hat{Z}} r(X)$.

Similarly, we should annex to $\hat{Z}$ any above average states in Z that are supported by $\hat{Z}$, namely,

$$\mathrm{free}(\hat{Z}) \equiv \text{largest set } V \subset Z \text{ such that:}$$

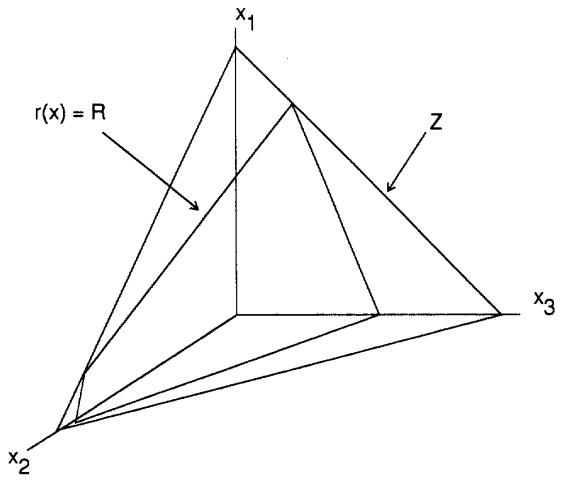


Fig. 3. The state space Z and the plane $r(x) = R$.

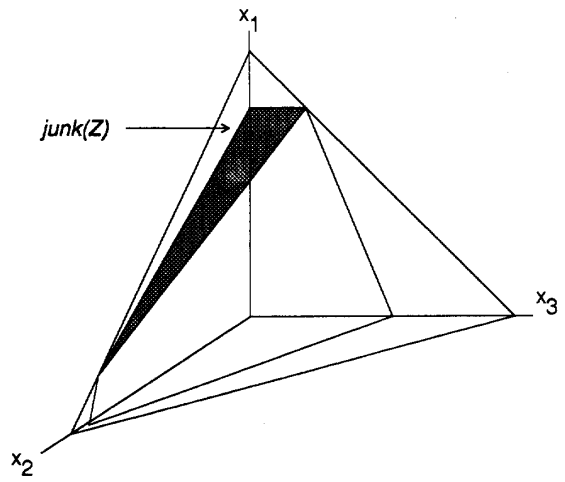


Fig. 4. Junk(Z).

$$r(x) > \hat{R} \qquad \forall x \in V$$

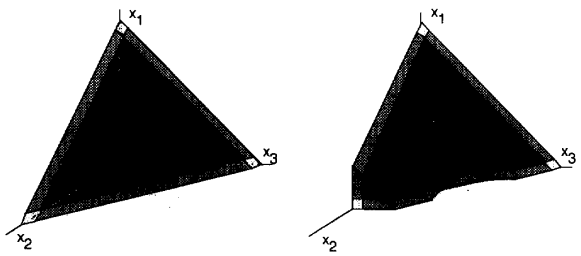
$$\& \, V \cap \hat{Z} = \varnothing$$

$$\& \, ss(V) \subseteq \hat{Z}$$

since $\mathrm{free}(\hat{Z})$ is annexable to $\hat{Z}$ and $E_{\mathrm{free}(\hat{Z})} r(X) > \hat{R}$.

To help picture what these sets may look like, consider a system with three service types and one resource type, as pictured in Fig. 3.

Junk(Z) is then those states in Z^- that we can remove without removing any states in Z^+, pictured as the set of states above the shaded triangle in Fig. 4. To see how we might use these constructs, we digress to a similar problem.



(a) *In the original state space* (b) *Cross sections displaying variation of r(x)*

Fig. 5. The rate of revenue earned in each state $r(x)$. (a) In the original state space. (b) Cross-sections displaying variation of $r(x)$.

IV. A MSSR Queuing System

Consider a system that can accept n service types but that can only work on one request of each type at a time. Assume the remainder of requests are accepted into the system and queued in a shared buffer if buffer space exists, and blocked if the buffer is full. Assume that each request queued occupies one slot in the shared buffer. Foschini and Gopinath [7] considered the problem of maximizing revenue in such a system and conjectured that the optimal coordinate convex policy is to limit sums of the number of each type in the system. We prove this conjecture here.

For a buffer with C slots, the state space would be

$$Z = \left\{ x \,\middle|\, \sum_i x_i \le C \right\}.$$

The transitions rates in this Markov chain are

$$r_{xy} = \begin{cases} \lambda_i, & \text{if } x \notin F_i \text{ and } y = (x_1, \cdots, x_i + 1, \cdots, x_n) \\ \mu_i, & \text{if } x \notin E_i \text{ and } y = (x_1, \cdots, x_i - 1, \cdots, x_n) \\ 0, & \text{else.} \end{cases}$$

The stationary distribution is product form [7]:

$$\pi(x) = \pi(0) \prod_{i=1}^{n} \rho_i^{x_i} \qquad \pi(0) = \frac{1}{\sum_{x \in Z} \prod_{i=1}^{n} \rho_i^{x_i}}.$$

The stationary distribution is time reversible, and hence coordinate convex control policies correspond to subsets of the state space with time reversible product form distributions, as in the MSMR model. For a system with three types of requests, $r(x)$ is pictured in Fig. 5(a), with darker shading representing higher $r(x)$. Cross sections are shown in Fig. 5(b) to help visualize the effect of placing restrictions upon the system. Let N be the set of all service types $\{1, \cdots, n\}$.

Theorem 1: There exist a set of constants[1] such that the optimal coordinate convex policy Z^* for the MSSR queueing system given above, can be represented as:

$$x \notin Z^* \quad \text{iff} \quad x \notin Z \quad \text{or}$$

$$\sum_{i \notin I} x_i > c_I \quad \text{for some} \quad I \subset N. \tag{2}$$

[1]For the remainder of this paper, $I \subset N$ is understood to exclude the case $I = \emptyset$.

or equivalently

$$x \in Z^* \quad \text{iff} \quad x \in Z \quad \text{and}$$

$$\sum_{i \notin I} x_i \le c_I \quad \text{for all} \quad I \subset N.$$

This form was called "simple" by Foschini and Gopinath.

Sketch of Proof: Define a projection function:

$$x \uparrow I \equiv (y_j), \qquad y_j = \begin{cases} x_j, & j \notin I \\ 0, & j \in I \end{cases} \quad \text{for } I \subset N.$$

An alternate representation to (2) is then

$$x \in Z^* \quad \text{iff} \quad x \in Z \quad \text{and} \quad x \uparrow I \in G_I^* \quad \text{for all } I,$$

$$\text{where} \quad G_I^* = \left\{ x \,\middle|\, \sum_{i \notin I} x_i \le c_I \right\}. \tag{3}$$

The proof proceeds by showing the following.

1) Z^* must contain all states in the interior of Z ($x_i > 0 \ \ \forall i$) that can be supported by Z^*, since these states are in free (Z^*). This reduces the problem of finding Z^* to that of finding sets G_I, with Z^* given by $x \in Z^*$ iff $x \in Z$ and $x \uparrow I \in G_I \ \ \forall I$.
2) The optimal G_I^* must be of the linear form of (3). This reduced the problem of finding Z^* to that of finding the constants $\{c_I\}$.

The full proof is presented in the Appendix.

This form for the optimal policy is very nice. Knowing that we can just look among policies that limit sums of services and that an optimal policy can be described by the set of constants $\{c_I\}$ is powerful information. For example, consider a two service type, one resource type queueing system. Theorem 1 guarantees that the optimal c.c. policy can be described by

$$Z^* = \{x \mid x \in Z, \quad x_1 \le c_1, \quad x_2 \le c_2\}$$

for some constants c_1 and c_2.

The question arises: is something similar true for the original MSMR model?

V. A Blocking MSMR System

A. The Optimal Coordinate Convex Policy

We first present a characterization of the optimal c.c. policy, and in the later sections get at what this means.

Theorem 2 : There exist a set of constants $\{c_I^k, I \subset N\}$ and a set of constants $\{\alpha_{i,k,I}\}$ such that the optimal coordinate convex policy Z^* for the MSMR model presented in Section II, can be represented as:

$$x \notin Z^* \quad \text{iff} \quad x \notin Z \quad \text{or} \quad \exists k \ni$$

$$\sum_{i \notin I} \alpha_{i,k,I} x_i > c_I^k \quad \text{for all} \quad I \subset N. \tag{4}$$

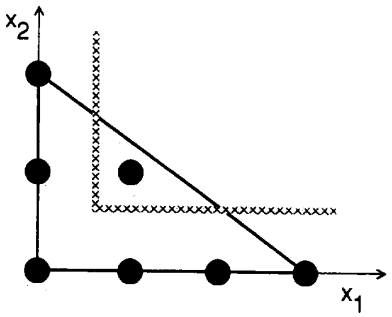


Fig. 6. A MSSR system with a nonconvex optimal c.c. policy.

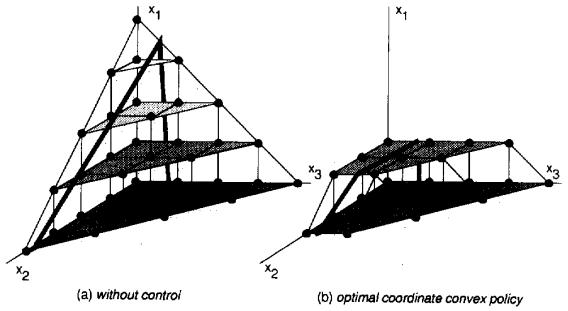


Fig. 7. A MSSR example. (a) Without control. (b) Optimal coordinate convex policy.

Sketch of Proof: The form of (4) is similar to that in Theorem 1 (2), except for two changes:

1) a removal of a set V from Z may now be characterized by the union of a *set* of inequalities rather than by a single inequality, due to lack of guaranteed convexity of Z^*. The k in (4) is used to label the set;
2) coefficients α are introduced to accommodate resource usage in quantities other than 1.

The proof proceeds by showing that if Z^* *cannot* be expressed in the claimed form, then there exists a set annexable to Z^* with average revenue greater than R^*, namely free(Z^*), or that there exists a set removable from Z^* with average revenue less than R^*, namely junk(Z^*).

The full proof is presented in the Appendix.

B. Commentary

Theorem 2 is a weaker characterization of the optimal policy than Theorem 1. This is due to a loss of guaranteed convexity of Z^*. To illustrate this, consider a system with two services and one resource type. Suppose service #1 requires two of the common resource, service #2 requires 3 and there are six resources in the system (Fig. 6).

Assume $r_1 = 2$, $r_2 = 3$, and that the load is very high. Then state $(1, 1)$ generates a rate of revenue $r(x) = 5$. The overall rate of revenue, however, is $R \cong 6$, since almost all the probability is in states $(0, 2)$ and $(3, 0)$. Thus, the optimal control policy would exclude state $(1, 1)$, namely, $Z^* = Z - \{(1, 1)\}$. Note that Z^* is *not* convex. The problem stems from resource usage in noninteger multiples.

If we restrict ourselves to *convex* policies, we can state a stronger characterization.

Corollary 1: There exist a set of constants $\{c_I, I \subset N\}$ and a set of constants $\{\alpha_{i,I}\}$ such that the optimal *convex* coordinate convex policy Z^* for the MSMR model presented in Section II, can be represented as

$$x \notin Z^* \quad \text{iff} \quad x \notin Z \quad \text{or}$$

$$\sum_{i \notin I} \alpha_{i,I} x_i > c_I \qquad \text{for some} \quad I \subset N \tag{5}$$

or equivalently

$$x \in Z^* \quad \text{iff} \quad x \in Z \quad \text{and}$$

$$\sum_{i \notin I} \alpha_{i,I} x_i \leq c_I \qquad \text{for all} \quad I \subset N.$$

For a two service one resource type system, Corollary 1 guarantees that the optimal *convex* c.c. policy can be described by

$$Z^* = \{x \mid x \in Z, \quad x_1 \leq c_1, \quad x_2 \leq c_2\}. \tag{6}$$

Convexity of Z^* holds for a few special MSSR systems.

First, if two services use a single common resource type, it has been shown that the optimal c.c. policy consists of a single threshold [24]. Second, if two services each share a common resource type and both

$$\frac{\#\text{ resources in system}}{\#\text{ resources used by service 2}}$$

and

$$\frac{\#\text{ resources used by service 2}}{\#\text{ resources used by service 1}}$$

are integers, it has been shown that the optimal c.c. policy consists of at most two thresholds as in (6) [24].

We will consider MSSR systems with greater than 2 services in a forthcoming paper [9].

C. Examples

Two examples are provided in this section.

First consider a MSSR system with three service types and one resource type, with each service requiring one of four available resources. The feasible state space Z is shown in Fig. 7(a). Assume that each service type has a load of 0.75, but that they pay differently, specifically that $r_1 = 1$, $r_2 = 2$, and $r_3 = 10$. Without any control, the system generates an average revenue of 8.56; the corresponding $r(x) = R$ plane is indicated by a dark band in Fig. 7(a).

The optimal coordinate convex policy, shown in Fig. 7(b), does not allow transitions to any state with $x_1 > 1$, or to states $(1, 3, 0)$ and $(0, 4, 0)$. It can be described by

$$Z^* = \{x \mid x \in Z, \quad x_1 \leq 1, \quad x_1 + x_2 \leq 3\}$$

and achieves a rate of revenue of 8.67.

In general, improvements to any policy are made by cutting off as much of the Z^- portion of the state space, and as little

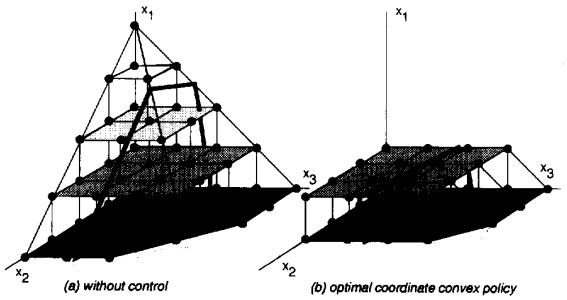


Fig. 8. A MSMR example. (a) Without control. (b) Optimal coordinate convex policy.

of the Z^+ portion, as possible without violating coordinate convexity. In this particular example, it is simple to conclude that eliminating state (4, 0, 0) is worthwhile, since it is in $Z^-[r(4, 0, 0) = 4 < 8.56]$ and since eliminating it does not violate coordinate convexity. It is also worth cutting off the entire third and second tiers, since here the Z^- portion outweights the Z^+ portion. However, at $x_1 = 1$, the Z^+ portion now outweighs the Z^- portion. Similarly, cuts are worthwhile starting in from the (0, 4, 0) corner $[r(0, 4, 0) = 8 < 8.56]$, in a slanted direction. The restriction of the state space increases the percentage of time that the system is close to the (0, 0, 4) corner, where it earns the highest rate of revenue $[r(0, 0, 4) = 40]$.

If the load were to increase, then the optimal coordinate convex policy would be constricted even more from that shown in Fig. 7(b). The $r(x) = R$ plane passes through the point EX, and has normal vector $(r_1, \cdots, r_n)$. Therefore, as the load increases, EX moves outward from the origin, the $r(x) = R$ plane moves outward from the origin, and more of the state space moves into Z^-. This means that deeper cuts become worthwhile.

As a second example, consider a MSMR system with three service types, and three types of resources. Assume that service type 1 requires one of resource type A and one of type B, service type 2 requires one of resource type A and one of type C, and service type 3 requires one of resource type B and one of type C. Assume that there are 4 resources of type A, 4 of type B, and 6 of type C in the system. The state space of this example is shown in Fig. 8(a). Assume that each service type has a load of 2, but that they pay differently, specifically that $r_1 = 1$, $r_2 = 1$, and $r_3 = 5$. Without any control, the system generates a average revenue of 9.86; the corresponding $r(x) = R$ plane is shown in Fig. 8(a).

As in the first example, the optimal coordinate convex policy cuts off portions of the state space near the relatively low paying corners (4, 0, 0) and (0, 4, 0) $[r(4, 0, 0) = r(0, 4, 0) = 4 < 9.86]$. This policy is shown in Fig. 8(b). It can be described by

$$Z^* = \{x \mid x \in Z, \quad x_1 \leq 1, \quad x_2 \leq 3\}$$

and generates a rate of revenue of 10.44.

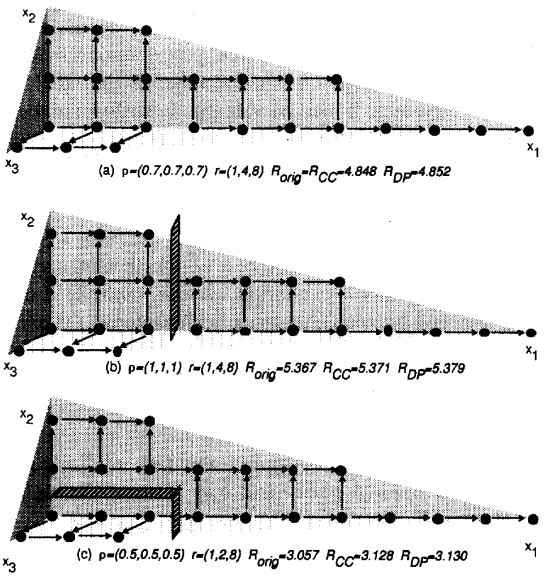


Fig. 9. Comparison of dynamic programming and coordinate convex solutions. (a) $\rho = (0.7, 0.7, 0.7) r = (1, 4, 8)$ $R_{\text{orig}} = R_{\text{CC}} = 4.848$ $R_{\text{DP}} = 4.852$. (b) $\rho = (1, 1, 1)$ $r = (1, 4, 8) R_{\text{orig}} = 5.367$ $R_{\text{CC}} = 5.371$ $R_{\text{DP}} = 5.379$. (c) $\rho = (0.5, 0.5, 0.5)$ $r = (1, 2, 8)$ $R_{\text{orig}} = 3.057$ $R_{\text{CC}} = 3.128$ $R_{\text{DP}} = 3.130$.

VI. Comparison to Dynamic Programming

We know, from an example presented by Ross and Tsang [24], that the optimal coordinate convex policy, which *restricts the state space*, sometimes produces a solution inferior to the optimal policy, found using dynamic programming, in a class of policies that *restrict individual transitions of the Markov chain*. In this section, we briefly compare these two different approaches using a few variations on Ross and Tsang's example.

Consider a MSSR system with 10 identical resources that accepts three types of requests. These service requests, if accepted, occupy 1, 4, and 8 resources, respectively. The state space is shown in each of Fig. 9(a)–(c). Detailed below each figure are the loads presented by each service type, and the rates of revenue paid by each. The optimal policy, found by dynamic programming, is indicated by arrows showing allowed arrivals. The optimal coordinate convex policy is given by the state space from the origin up to striped bars, if any. The resulting average rates of revenue for the uncontrolled (R), coordinate convex (R_{CC}), and dynamic programming (R_{DP}) solutions are shown below each figure.

It is evident that the two types of control can give very different solutions. The dynamic programming technique tends to break links one at a time, as the load increases, until the state space becomes reducible. The coordinate convex policy, unable to exercise this type of fine control, generates this reduced state space at a load lower than that at which the dynamic programming solution reduces, but higher than that at which the dynamic programming solution starts to deny arrivals out of the subset.

It seems, however, that the two solutions, although sometimes very different, generate average rates of revenue that are not very different. Since the dynamic programming technique quickly becomes burdensome to calculate as the size of the state space increases, it may be valuable to write an algorithm to find the optimal coordinate convex policy, or a close approximation to that policy.

VII. Parting Thoughts

We have considered the decision of whether to accept or deny service requests in a multiple service, multiple resource system. We have been able to strongly characterize the optimal policy for a simpler but related queueing system, and weakly characterize the optimal policy for our MSMR system, by restricting ourselves to policies that base this decision only on the state the system would enter if the request were granted.

These characterizations simplify the task of finding and describing the optimal policy and may provide the basis for more reasonable algorithms than have yet been devised.

Future work will address the descriptive gap between the two characterizations presented here, and the lack of computationally feasible approaches to designing management strategies for particular MSMR systems.

Appendix
Proofs

A. Theorem 1

Define the marginal expectation of X_i, conditioned on a maximum value

$$m_i(y) \equiv E_{f_i}(X_i \mid X_i \le y) = \frac{\sum_{x=0}^{y} x f_i(x)}{\sum_{x=0}^{y} f_i(x)} \qquad \text{where}$$

$$f_i(x) = \prod_{i=1}^{n} \rho_i^{x_i}$$

and define a cross section:

$$C_Z(x \uparrow I) = \{y \in Z \mid y \uparrow I = x \uparrow I\}.$$

Note that $C_Z(x \uparrow I)$ is a $|I|$ dimensional subset of Z. And finally, define a translation operator:

$$T_{jk}(x) = (y_i), \qquad y_i = \begin{cases} x_i, & i \neq j, k \\ x_i - 1, & i = j \\ x_i + 1, & i = k \end{cases}.$$

We will use $T_{jk}(x)$ to take advantage of the uniform one buffer slot per queued request.

First we will show that Z^* must contain all states in the interior of Z ($x_i > 0\ \forall i$) that can be supported by Z^*. Focus on the set of states that have been removed from Z to obtain Z^*, namely,

$$V \equiv Z - Z^*$$

Define W to be those states in V, if any, that are in the interior of Z and can be supported by Z^*, namely,

$$W \equiv \{x \in V \mid x \uparrow \{i\} \in Z^* \quad \forall i\}.$$

Now

$$x \in W \Rightarrow r(x) = \sum_i r_i$$

which is the highest rate of revenue, so certainly $E_W r(X) > R^*$. By construction, W is annexable to Z^* since $ss(W) \in Z^*$. Therefore, we can add W to Z^* and it would increase the average revenue to do so. Thus, if Z^* is optimal, $W = \emptyset$.

This reduces the problem of finding Z^* to that of finding sets G_I, with Z^* given by $x \in Z^*$ iff $x \in Z$ and $x \uparrow I \in G_I \quad \forall I$. It remains to show that the optimal sets G_I^* are given by the G_I^* in (3). Partition V into sets U_I, $I \subset N$ where $x \in U_I \Leftrightarrow x \uparrow I \notin Z^*$ and $x \uparrow \{i\} \in Z^*$, $i \notin I$. The set I will correspond to the subscripts $\{i\}$ which don't appear in the I constraint.

Fix I, and focus upon the states U_I removed by the G_I constraint. Fix some $x \in U_I$. Since all queued requests occupy *one* buffer space, exchanging one of service j for one of service k does not affect the capacity for services other than j or k [See Fig. 5(b)]. Formally, we can express this as: the values of the "I" components of the "I" type cross sections are invariant with respect to $T_{j,k}$ operations ($j, k \notin I$), namely,

$$T_{j,k}[C_Z(x \uparrow I)] = C_Z(T_{j,k}[x] \uparrow I)$$
$$j, k \notin I \text{ (when the quantities are defined).}$$

Furthermore, the distribution on Z is product form, hence the distribution (of the "I" components of x) on these cross sections, namely, $\text{dist}[x \uparrow (N - I) \mid x \in C_Z(x \uparrow I)]$, are invariant with respect to $T_{j,k}$ operations.

Therefore, certainly the *expectation* on these cross sections, $E_{C(x\uparrow I)} r(X)$, is constant with respect to $T_{j,k}$ operations on $x \uparrow I$, ($j, k \notin I$). We can conclude that if Z^* is optimal, then cuts of type I must be made with slope 1, namely, $\boldsymbol{x \notin Z^*}$ $(\& \ x \in Z) \Rightarrow \boldsymbol{T_{j,k}(x) \notin Z^*}$. Therefore the upper boundaries of the optimal G_I^* are of slope 1 and hence given by

$$G_I^* = \left\{ x \;\middle|\; \sum_{i \notin I} x_i \le c_I \right\}.$$

This concludes the proof.

B. Theorem 2

The discreteness and coordinate convexity of the state space implies that any constriction of the state space can be described by a set of hyper-planar faces, with states on one side of each face to remain in the optimal state space, and states on the other side of the face to be excluded from the optimal state space. In a discrete space, however, many hyperplanes may accomplish the same division of states; therefore, in the remainder of this paper, we uniquely define the hyperplane representing a face to be the one which passes through the states nearest the division that are to be included in the optimal state space. Note this "connect-the-dots" approach implies that each $\alpha_{i,k,I} \ge 0$.

The proof proceeds by showing that if Z^* is not the claimed form, then $\exists \text{junk}(Z^*) \neq \emptyset$ or $\exists \text{free}(Z^*) \neq \emptyset$, and that $\hat{Z} = Z^* - \text{junk}(Z^*) + \text{free}(Z^*)$ generates a higher rate of revenue than Z^*. Suppose $Z^* \subseteq Z$ is optimal. We claim that Z^* can be represented in the form of (4). Now Z^* can be

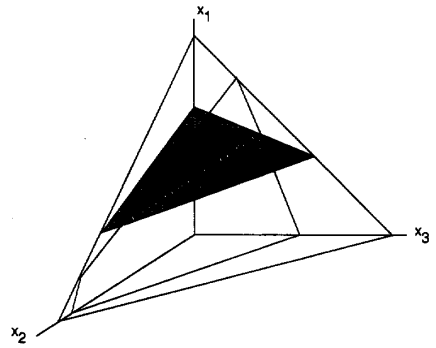


Fig. 10. A proposed removal.

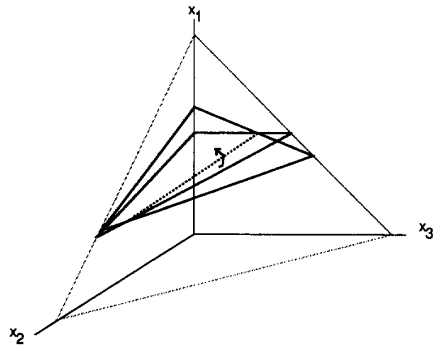


Fig. 11. Rotating the face to produce a better face.

represented as the result of a series of exclusions of removable sets, so it suffices to show that each removal can be represented by a set k of equations in the form of (4).

Focus on one such removal, of a set V from Z' to obtain Z'' where $Z \supseteq Z' \supset Z'' \supseteq Z^*$. Now if $E_V r(x) < R'$ then V must either intersect the $r(x) = R'$ hyperplane, or lie entirely below it. We consider these two cases separately. First consider the case where V lies entirely below the $r(x) = R'$ hyperplane. Since $V \subset Z'^-$ and V is removable from Z', it follows that $V \subseteq \text{junk}(Z')$. By construction, however, $\text{junk}(Z')$ can be represented by a set k of equations in the form of (4). Therefore, if Z^* is optimal, V can be represented by a set k of equations in the form of (4).

Now consider the case where V intersects the $r(x) = R'$ hyperplane. For the example in Figs. 3–4, such a set is shown in Fig. 10 as the region above the shaded triangle.

Consider a related set V^*, obtained as follows: rotate each face of V that intersects $r(x) = R'$ until one of the coefficients of the face is zero and the rest are positive. V^* is also removable from Z'. (Rotating a face to obtain a *negative* coefficient would have invalidated this claim.) Furthermore, V^* can be represented as a set k of equations in the form of (4).

Now $V - V \cap V^* \subseteq \text{junk}(Z')$ and $V^* - V \cap V^* \subseteq \text{free}(Z')$. Therefore, V^* is at least as good a removal as V. This if Z^* is optimal, V can be represented by a set k of equations in the form of (4).

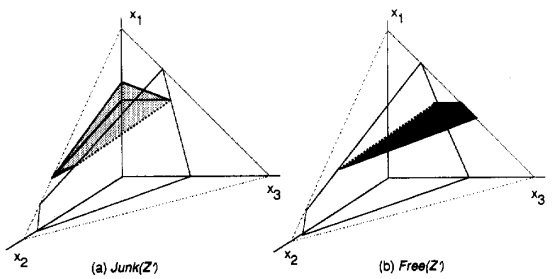


Fig. 12. The difference between the original and altered faces. (a) Junk (Z'). (b) Free (Z').

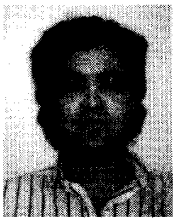
For our running example, this alteration is shown in Figs. 11 and 12. Note the intersection of the $r(x) = R'$ plane and the proposed face, shown as a dashed line. Junk (Z') and free (Z') can be thought of as the product of rotating the face about this intersection until it is constant with respect to at least one service type.

This concludes the proof.

REFERENCES

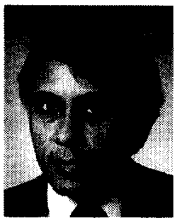
[1] J. M. Aein, "A multi-user-class, blocked-calls-cleared, demand access model," *IEEE Trans. Commun.*, vol. COM-26, pp. 378–385, Mar. 1978.
[2] F. Baskett, K. M. Chandy, R. R. Muntz, and F. G. Palacios, "Open, closed, and mixed networks of queues with different classes of customers," *J. Assoc. Comput. Machinery*, vol. 22, no. 2, pp. 248–260, Apr. 1975.
[3] P. Bloom and P. Miller, "Intelligent network/2," *Telecommun. Int.*, Feb. 1987.
[4] D. Y. Burman, J. P. Lehoczky, and Y. Lim, "Insensitivity of blocking probabilities in a circuit-switched network," *J. Appl. Prob.*, vol. 21, pp. 850–859, 1984.
[5] Z. Dziong and J. W. Roberts, "Congestion probabilities in a circuit-switched integrated services network," *Perform. Eval.*, vol. 7, pp. 267–284, 1987.
[6] G. J. Foschini, B. Gopinath, and J. F. Hayes, "Optimal allocation of servers to two types of competing customers," *IEEE Trans. Commun.*, vol. 29, pp. 1051–1055, July 1981.
[7] G. J. Foschini and B. Gopinath, "Sharing memory optimally," *IEEE Trans. Commun.*, vol. 31, pp. 352–360, Mar. 1983.
[8] S. Jordan and P. P. Varaiya, "Throughput in multiple service, multiple resource communications networks," *IEEE Trans. Commun.*, vol. 39, pp. 1216–1222, Aug. 1991.
[9] S. Jordan, "A continuous state space model of multiple service, multiple resource communication networks," to appear in *IEEE Trans. Commun.*
[10] J. S. Kaufman, "Blocking in a shared resource environment," *IEEE Trans. Commun.*, vol. 29, pp. 1474–1481, Oct. 1981.
[11] F. P. Kelly, "Networks of queues with customers of different types," *J. Appl. Prob.*, vol. 12, pp. 542–554, 1975.
[12] ——, "Networks of queues," *Advances Appl. Prob.*, vol. 8, pp. 416–432, 1976.
[13] ——, *Reversibility and Stochastic Networks.* New York: Wiley, 1979.
[14] ——, "Blocking probabilities in large circuit-switched networks," *Advances Appl. Prob.*, vol. 18, pp. 473–505, 1986.
[15] ——, "Routing in circuit-switched networks: Optimization, shadow prices and decentralization," *Advances Appl. Prob.*, vol. 20, pp. 112–144, 1988.
[16] B. Kraimeche and M. Schwartz, "Circuit access control strategies in integrated digital networks," in *Proc. IEEE Conf. Inform. Syst.*, 1984, pp. 320–235.
[17] S. S. Lam, "Queueing networks with population size constraints," *IBM J. Res. Develop.*, pp. 370–378, July 1977.
[18] L. Ludwig, "A threaded/flow approach to reconfigurable distributed systems and service primitive architectures," *ACM SigCom*, Stowe, VT, Aug. 1987.
[19] D. Mitra, "Asymptotic analysis and computational methods for a class of simple, circuit-switched networks with blocking," *Advances Appl. Prob.*, vol. 19, pp. 219–239, 1987.

[20] P. Nain, "Qualitative properties of the Erlang blocking model with heterogeneous user requirements," *Queueing Syst.: Theory and Appl.*, vol. 6, pp. 189–206, 1990.
[21] L. Pate, "IMAL analytical study: the interconnection of switches and concentrators with shared resources," *Bell Commun. Res.*, preprint, Mar. 1989.
[22] S. Ross, *Stochastic Processes.* New York: Wiley, 1983.
[23] K. W. Ross and D. Tsang, "Optimal circuit access policies in an ISDN environment: A Markov decision approach," *IEEE Trans. Commun.*, vol. 37, pp . 934–939, Sept. 1989.
[24] ——, "The stochastic knapsack problem," *IEEE Trans. Commun.*, vol. 37, pp. 740–747, July 1989.
[25] ——, "Teletraffic engineering for product-form circuit-switched integrated services networks," *Advances Appl. Prob.*, vol. 22, pp. 657–675, 1990.
[26] K. W. Ross and D. D. Yao, "Monotonicity properties for the stochastic knapsack," *IEEE Trans. Inform. Theory*, vol. 36, pp. 1173–1179, 1990.
[27] E. Souza e Silva and R. R. Muntz, "Simple relationships among moments of queue lengths in product form queueing networks," *IEEE Trans. Comput.*, vol. 37, pp. 1125–1129, Sept. 1988.
[28] J. T. Virtamo, "Reciprocity of blocking probabilities in multiservice loss systems," *IEEE Trans. Commun.*, vol. 36, pp. 1174–1175, Oct. 1988.
[29] H. Watanabe, "Integrated office systems: 1995 and beyond," *IEEE Commun. Mag.*, vol. 25, pp. 74–78, Dec. 1987.
[30] S. B. Weinstein, "Telecommunications in the coming decade," *IEEE Spectrum*, pp. 62–67, Nov. 1987.
[31] W. Whitt, "Blocking when service is required from several facilities simultaneously," *AT&T Tech. J.*, vol. 64, pp. 1807–1856, 1985.
[32] *Digest of Intelligent Networks Workshop*, Lake Yamanaka, Japan, Oct. 1989.
[33] S. Zachary, "Control of stochastic loss networks, with applications," *J. Roy. Statist. Soc. B*, vol. 50, no. 1, pp. 61–73, 1988.
[34] L. Zhang, "Designing a new architecture for packet-switching communication networks," *IEEE Commun. Mag.*, vol. 25, pp. 5–12, Sept. 1987.



Scott Jordan (S'86–M'90) received the B.S./A.B., M.S., and Ph.D. degrees from the University of California, Berkeley, in 1985, 1987, and 1990, respectively.

He is currently an Assistant Professor at Northwestern University, Evanston, IL. His teaching and research interests are the modeling and analysis of behavior, control, and pricing of computer/telecommunication networks.



Pravin P. Varaiya (M'68–SM'78–F'80) received the Ph.D. degree in electrical engineering from the University of California, Berkeley.

He is currently Professor of Electrical Engineering and Computer Sciences, University of California, Berkeley. He is the coauthor, along with P. R. Kumar, of *Stochastic Systems: Estimation, Identification, and Adaptive Control* (Englewood Cliffs, NJ: Prentice-Hall), 1986, and coeditor, with A. Kurzhanski, of *Discrete Event Systems: Models and Applications*, Lecture Notes in Information Sciences (New York: Springer), vol. 103, 1988. His areas of research and teaching are in stochastic systems, communication networks, power systems, and transportation.