Title

Throughput in multiple service, multiple resource communication networks

Journal
IEEE Transactions on Communications, 39(8)

Authors
Jordan, Scott
Varaiya, Pravin

Publication Date
1991-08-01

DOI
10.1109/26.134011

Peer reviewed

Throughput in Multiple Service, Multiple Resource Communication Networks

Scott Jordan 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 pose resource allocation problems, present a sensitivity analysis, and provide a glimpse of the possible behavior of such networks.**

The simplest discipline is assumed: a service request is accepted if the necessary resources are available; otherwise it is rejected. Two results are obtained. The first gives the sensitivity of throughput of service requests of type *i* with respect to offered traffic and service rates of type *j*. The second result is that the set of vectors of achievable throughput rates is a convex polyhedron given by an explicit set of linear inequalities.

I. Introduction

In this paper, we focus upon communication networks that integrate multiple services using multiple resources. In particular, we pose resource allocation problems, present a sensitivity analysis, and provide a glimpse of the possible behavior of such networks.

This work is motivated by several trends in networks. 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.

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.

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 decisions 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 this paper, we investigate the nature of this interaction. In future papers, we will address issues of control and pricing of such a system.

Considerable effort has been put into understanding related but simpler multiple service, *single resource* (MSSR) systems. In [1], Aein constructed a Markov chain model and stated the resulting product form stationary distribution. Kaufman [2] showed that this product form holds under more general assumptions, including general service distributions. More recent papers exhibit the relationship between traffic intensity and throughput: Virtamo [4] displays a reciprocity relation in the sensitivity of blocking probabilities to traffic intensity and Ross and Yao [9] and Nain [10] investigate the effect of increasing traffic intensity upon throughput.

Some effort has also been applied to MSMR systems. In [12], [13], 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 [14], Burman et al obtain an insensitivity result for the stationary distribution of a MSMR system. Numerical aspects have been investigated in [15]–[18].

In addition, the MSMR system considered here is similar to some queueing systems. Foschini and Gopinath [3] investigated control policies to maximize throughput or minimize blocking probabilities in a MSSR queueing system. E. Souza, E. Silva, and Muntz [23] have recently displayed sensitivity

Paper approved by the Editor for Queueing and Networking Performance of the IEEE Communications Society. Manuscript received April 12, 1988; revised October 18, 1989. 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 supported by the Fannie and John Hertz Foundation. This paper was presented at the Conference on Decision and Control, Tampa, FL, December 1989.

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

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

IEEE Log Number 9100830.

results for product form queueing systems.

The MSMR model is investigated here for the simplest discipline: a request is granted if the necessary servers are available; otherwise it is rejected. Section II displays the model. In Section III, we present sensitivity results and discuss the implications of these upon the nature of multiple service communication networks. In Section IV, we study the range of achievable throughput rates. In Section V, we consider relaxations of the statistical assumptions. Some closing comments are in Section VI, and some proofs are in the appendix.

II. 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 processing starts immediately; if the necessary resources are unavailable then the request is lost to the system.

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 and adopt the following notation.

$\lambda \equiv (\lambda_1, \cdots, \lambda_n)$, 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.

$Z \equiv \{x | x$ is feasible, i.e., x can be simultaneously processed with the available resources$\}$.

$F_i \equiv \{x \,|\, x \in Z$ but $(x_1, \cdots, x_i + 1, \cdots, x_n) \notin Z\}$.

$E_i \equiv \{x \,|\, x \in Z$ but $(x_1, \cdots, x_i - 1, \cdots, x_n) \notin Z\}$.

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

$r_{xy} \equiv$ rate of transitions to state y, given we are currently in state x.

$P(\cdot) \equiv$ Probability of $\cdot$.

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

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

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 $(x_1, \cdots, x_i - 1, \cdots, x_n) \in Z$.

As an example, consider a system that accepts only two types of requests: type 1 requires one of resource A and one of resource B, and request 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.

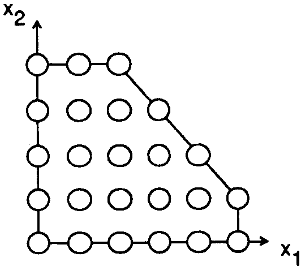


Fig. 1. The state space Z for a two service type, three resource type system.

The key is to model the *services* directly, rather than to model the resource space. Resource constraints thus appear indirectly, in the shape of the state space Z, as constraints upon the number of each type of service that can be provided simultaneously.

The Markov chain is time reversible. The local balance equations are

$$\pi(x_1, \cdots, x_i, \cdots, x_n)x_i\mu_i = \pi(x_1, \cdots, x_i - 1, \cdots, x_n)\lambda_i \quad \forall\, x \in Z \quad \ni x_i \geq 1.$$

Conservation of Probability implies $\sum_{x \in Z} \pi(x) = 1$.

Iterating the balance equations yields the well known product form stationary distribution

$$\pi(x) = \pi(0) \prod_{i=1}^{n} \frac{\rho_i^{x_i}}{x_i!} \quad \text{where } \pi(0) \equiv \pi(0, \cdots, 0). \tag{1}$$

Conservation of Probability gives us the normalization constant

$$\pi(0) = \frac{1}{\sum_{x \in Z} \prod_{i=1}^{n} \frac{\rho_i^{x_i}}{x_i!}}. \tag{2}$$

III. Sensitivity Results

In this section, we investigate the sensitivities of the throughput rates L_i and the blocking probabilities $P(F_i)$ to the request rates λ_i the service rates μ_i and the loads, ρ_i.

Theorem 1:

$$\frac{\partial L_i}{\partial \lambda_j} = \begin{cases} \frac{\mu_i}{\lambda_j} \operatorname{cov}(x_i, x_j), & \text{if } i \neq j \\ \frac{\mu_i}{\lambda_i} \operatorname{var}(x_i), & \text{if } i = j \end{cases}. \tag{3}$$

The proof in the appendix relies on Little's result which, applied to this system, gives

E(number of type i in the system)

= (Average arrival rate into the system)

$\cdot$ (Average length of time in the system)

$$\text{i.e. } E(x_i) = L_i(1/\mu_i). \tag{4}$$

Using the stationary distribution (1) and (2) above, differentiating (4) with respect to λ_j, and transforming the differentiation operation into an expectation operation produces (3).

Similarly, we can find the following sensitivities:

$$\frac{\partial L_i}{\partial \mu_j} = \begin{cases} -\frac{\mu_i}{\mu_j}\,\mathrm{cov}(x_i, x_j), & \text{if } i \neq j \\ E(x_i) - \mathrm{var}(x_i), & \text{if } i = j \end{cases}$$

$$\frac{\partial \boldsymbol{P}(F_i)}{\partial \lambda_j} = \begin{cases} -\frac{\mu_i}{\lambda_i \lambda_j}\,\mathrm{cov}(x_i, x_j), & \text{if } i \neq j \\ \frac{\mu_i}{\lambda_i^2}\,[E(x_i) - \mathrm{var}(x_i)], & \text{if } i = j \end{cases}$$

$$\frac{\partial \boldsymbol{P}(F_i)}{\partial \mu_j} = \begin{cases} \frac{\mu_i}{\lambda_i \mu_j}\,\mathrm{cov}(x_i, x_j), & \text{if } i \neq j \\ -\frac{1}{\lambda_i}\,[E(x_i) - \mathrm{var}(x_i)], & \text{if } i = j \end{cases}$$

$$\frac{\partial \boldsymbol{P}(F_i)}{\partial \rho_j} = \begin{cases} -\frac{\mu_i \mu_j}{\lambda_i \lambda_j}\,\mathrm{cov}(x_i, x_j), & \text{if } i \neq j \\ \frac{\mu_i^2}{\lambda_i^2}\,[E(x_i) - \mathrm{var}(x_i)], & \text{if } i = j \end{cases}$$

$$\frac{\partial E(x_i)}{\partial \rho_i} = \frac{1}{\rho_i}\,\mathrm{var}(x_i)$$

$$\frac{\partial\, \mathrm{var}(x_i)}{\partial \rho_i} = \frac{1}{\rho_i} E(x_i - E(x_i))^3.$$

In particular, Virtamo's [4] reciprocity relation $\partial \boldsymbol{P}(F_i)/\partial \rho_j = \partial \boldsymbol{P}(F_j)/\partial \rho_i$ follows from the equation for $\partial \boldsymbol{P}(F_i)/\partial \rho_j$ above.

We study these sensitivity results for the case $i \neq j$ in Section III-A, and for the case $i = j$ in Section III-B.

A. Cross Sensitivities

The signs of *cross* sensitivities all depend on the sign of the associated $\mathrm{cov}(x_i, x_j)$, which in turn depends on the variation of $E(x_j \mid x_i)$ with respect to x_i. If $E(x_j \mid x_i)$ increases with x_i, the covariance is positive. By (3), this implies $\partial L_i/\partial \lambda_j > 0$, indicating any increase in the rate of type j service requests actually *increases* the throughput of type i (and vice versa). If this is true, we say that these two services are *complements.* Similarly, if $E(x_j \mid x_i)$ decreases with x_i, the covariance is negative and these two services are *substitutes.* If the variation of $E(x_j \mid x_i)$ is not monotonic, then the sign of the covariance is not so easily determined.

A few examples help to illustrate this. First consider a system with three service types 1, 2, and 3. Suppose that service type 1 requires one of resource A, service type 2 requires one C, and service type 3 requires one A and one C. The state space is pictured in Fig. 2(a); it is drawn as a continuous region for easier conceptualization. Simple analysis shows that $E(x_2 \mid x_1)$ increases with x_1, and that $E(x_3 \mid x_1)$ decreases with x_1; hence, services 1 and 2 are complements while services 1 and 3 are substitutes. This should be no surprise. Services 1 and 3 compete for resource A, so increasing the rate of type 1 service requests decreases the throughput of type 3; this blocking of type 3 also increases the throughput of type 2.

As a second example, consider the same system but now suppose service types 1 and 2 also use one of resource B each, as pictured in Fig. 2(b). All services are now substitutes.

As a final example, consider the same system as in the second example but now suppose that the number of available B's is higher. The new state space is pictured in Fig. 2(c). We find that $E(x_3 \mid x_1)$ decreases with x_1, but $E(x_2 \mid x_1)$ first increases and later decreases as x_1 increases. Thus, service types 1 and 3 are still substitutes, but we cannot conclude anything about the relationship between service types 1 and 2.

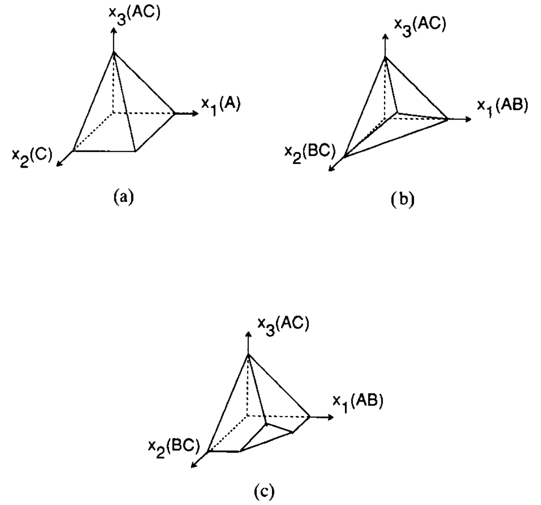


Fig. 2. Examples of substitutes and complements.

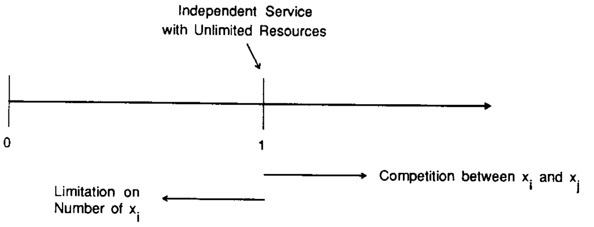


Fig. 3. Factors affecting the ratio of variance to mean of x_i.

B. Self-Sensitivities

From Theorem 1, we know that increasing the arrival rate of one type of service request always *increases* the rate at which that service type is accepted into the system, i.e., $\partial L_i/\partial \lambda_i > 0$. However, the sign of all the other self-sensitivities depend on the ratio of the variance to the mean of the x_i distribution. Thus, for instance, increasing the service rate for one type does *not* always increase the rate at which that service type is accepted into the system.

If $Ex_i > \mathrm{var}(x_i)$, then $\partial L_i/\partial \mu_i > 0$; we say that service type i is *self-advantageous.* Similarly, if $Ex_i < \mathrm{var}(x_i)$, then $\partial L_i/\partial \mu_i < 0$; we say that service type i is *self-disadvantageous.* We investigate the sign of $Ex_i - \mathrm{var}(x_i)$ by looking at the ratio $\mathrm{var}(x_i)/Ex_i$, and we note that it varies with μ_i. Some factors affecting this ratio are shown in Fig. 3.

Some examples help illustrate these factors. First, consider a system with just one service type. Assume that up to N of this service can be provided simultaneously. The resulting Markov chain is pictured in Fig. 4(a); we have not labeled the transitions to increase clarity. If $\mu_1 = 0$, then $x_1 = N$ almost surely, and accordingly the ratio of the variance to the mean is 0. As μ_1 increases, the effect of the barrier at $x_1 = N$ lessens, and the ratio increases. As μ_1 approaches infinity, $E(x_1)$ drops toward 0, but the distribution of x_1 tends toward a Poisson

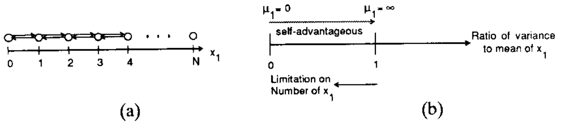

0
1
2
3
4
N
x_1
$\mu_1 = 0$
$\mu_1 = \infty$
self-advantageous
Ratio of variance to mean of x_1
0
1
Limitation on Number of x_1
(a)
(b)

Fig. 4. An example of a self-advantageous service.

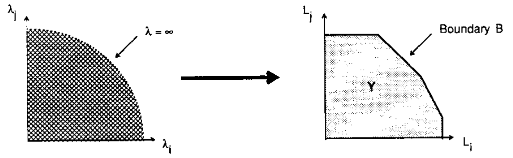


Fig. 6. Mapping the λ space into the L space.

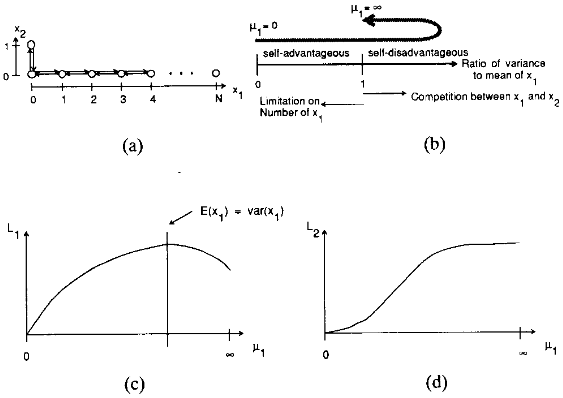


Fig. 5. An example of a self-disadvantageous service.

distribution, and the ratio of the variance to mean approaches 1 accordingly. This tendency is shown in Fig. 4(b). Note that for any value of μ_1, service type 1 is self-advantageous.

Now consider the same system, but with a second service type. Assume that only one of type 2 can be provided at a time, and only if none of type 1 are in the system. The resulting Markov chain is pictured in Fig. 5(a). As in the first example, if $\mu_1 = 0$, then $x_1 = N$ almost surely, and the ratio of the variance of x_1 to its mean is 0. As μ_1 increases, the effect of the barrier at $x_1 = N$ lessens, and competition between x_1 and x_2 increases, and thus the ratio increases, eventually pushing past 1. As μ_1 approaches infinity, $E(x_1)$ drops toward 0, and competition between x_1 and x_2 decreases as the system becomes mostly idle; the ratio of the variance to mean again approaches 1. This tendency is shown in Fig. 5(b).

So increasing the type 1 service rate increases the rate of acceptance for type 1 only until $E(x_1) = \text{var}(x_1)$; after that point L_1 decreases with increasing μ_1. As pictured in Fig. 5(c)–(d), service types 1 and 2 are always substitutes, but service type 1 changes from self-advantageous to self-disadvantageous.

C. Discussion

In summary, Theorem 1 provides three results. First, it lends a characterization of pairs of services as complements, substitutes, or both, depending upon the sign of the associated covariance of the number of each in the system. Second, it states that increasing the arrival rate of one request type always increases the rate at which that type is accepted into the system. Third, it lends a characterization of each single service, given a specified set of arrival and service rates, as advantageous or disadvantageous, depending upon the ratio of the variance to expectation of the number in the system.

These results can be specialized to a single resource (MSSR) model $n = 1$ to obtain some of the conclusions reached in [9], [10]. Consider links between two nodes as the single resource, and requests for bandwidth as multiple services (with the service type given by the number of links required). Using the first result, conclusions such as "requests for the largest bandwidth permissible compete with all other requests" can be made, permitting that the discrete state space satisfies certain regularity conditions (see [9]). The second result, applied to this system states "the throughput of requests for any particular bandwidth is an increasing function of the rate of such requests." The third result states that throughput is not necessarily monotonically increasing in the service rate (see [10]).

These results have particular relevance to the design of communication systems. Theorem 1 provides help when sizing a communication system, or when estimating the impact of a new service offering on existing services. Applied to simulations, sensitivities may be calculated from covariances obtained without having to perturb parameters.

IV. Achievable Throughput

In this section, we look at the region of achievable throughput. The set of all possible arrival rates $\lambda \in \mathcal{R}_+^n$ maps via (1), (2), and (4) into some region Y in the L space. We investigate the shape of Y. A two-dimensional slice of this is shown in Fig. 6.

Theorem 2: Y is a convex polyhedron. Moreover, if $\{x^*\}$ are the extreme points[1] of the state space Z, then $\{\mu x^*\}$ are the vertices of Y.[2]

The proof in the Appendix proceeds by showing that the infinite boundary in the λ space maps into the boundary B of Y; that if $\lambda = \infty$, then all states with nonzero probability lie on a hyperplane tangent to the state space Z from above; that this implies that L must lie on an equivalent hyperplane in its space; and that therefore the region Y is the convex hull formed by these hyperplanes. It relies in part on Theorem 1, and especially on the linear relationship between x_i and L_i expressed in (4).

Theorem 2 suggests a revenue optimization problem. Suppose that each service performed generates a revenue of $\$r_i$.

[1] $x^* \in Z$ is an extreme point of Z if x^* cannot be expressed as a convex combination of other states in Z.

[2] The multiplication μx^* is taken componentwise, i.e., $\mu x^* = (\mu_1 x_1^*, \cdots, \mu_n x_n^*)$.

Then maximization of the total revenue

$$\sum_{x \in Z} \sum_{i=1}^{N} r_i \lambda_i 1_{(x \notin F_i)} = \sum_{i=1}^{N} r_i L_i$$

corresponds to maximizing this linear function over Y. This is a linear programming problem.

V. Extensions

In this section, we consider the effects of relaxations of the statistical assumptions of the model posed in Section II upon Theorems 1 and 2.

First, consider arbitrary service distributions, with rational Laplace transforms, with means $1/\mu_i$. In [14], Burman et al show that the distribution (1–2) is insensitive to the service distribution. Little's result, $E(x_i) = L_i(1/\mu_i)$, also still holds, with $1/\mu_i$ now interpreted as the mean service time for service type i. The proofs of Theorems 1 and 2 thus follow as before.

Second, consider noncoordinate convex sample spaces. This might occur if service completion requires some other event to occur first, or if service completion requires another service to start, and this new service is blocked. Theorem 2 will hold provided that departures are never blocked while the system is in any of the states on the upper boundary $(\cup F_i)$ of the state space Z, but Theorem 1 would require consideration of state dependent service rates, since the average length of time in the system for a service of type i is no longer $1/\mu_i$.

Third, consider state dependent arrival rates. This may occur if the queueing system is closed (e.g., see [9]). Theorems 1 and 2 hold if $\lambda_i(x) = f(x_i)\lambda_i$ where λ_i is a constant. (See [11] and [24] for background on truncated multidimensional birth–death processes.)

Finally, consider state dependent departure rates. This allows some alternative service disciplines, e.g., processor sharing; see [5]–[7] for more detail on alternatives. It does *not,* however, allow for general queueing schemes; general queueing would produce departure rates that depend not only upon state but also upon the path to that state i.e. upon which how many of each service type were queued and how many were in service. Little's formula now becomes $E(x_i) = L_i E(1/\mu_i)$, and we lose the linear relationship between x_i and L_i. Theorems 1 and 2 no longer hold. (See [11] and [24].)

VI. Conclusion

We have analyzed the simplest MSMR model of a communications system which can process general types of requests, each of which requires several types of resources. More realistic models will have to abandon two assumptions.

First, we assumed that the resources needed to process a service request are acquired and released simultaneously. In practice the situation is more complex. For instance, in processing a credit card call, a database query is first made to verify the status of the caller; after it is approved the call is processed. Thus the two resources—database transaction and call handling—are occupied sequentially. On the other hand, in a conference call several links are occupied concurrently. In general, then, processing a service request can require a combination of sequential and concurrent access to resources. We need new approaches to specify such service requests and to model the scheduling of resources [21].

Second, we assumed the simplest discipline. Two extensions are worth considering: requests can be queued, and resources can be reserved to ensure fairness and in anticipation of future revenue generating requests [22].

Appendix

A. Proof of Theorem 1

Proof of Theorem 1: Since service requests are accepted whenever feasible, L_i can be related to λ_i by

$$L_i = \lambda_i[1 - \boldsymbol{P}(F_i)]. \tag{5}$$

A better relation to start with, however, comes from viewing this as a queueing system for service type i and using Little's result:

$$\begin{aligned} &E(\text{number of type } i \text{ in the system}) \\ &\qquad = (\text{Average arrival rate in the system}) \\ &\qquad \cdot (\text{Average length of time in the system}) \\ &\qquad\qquad \text{i.e., } E(x_i) = L_i(1/\mu_i). \end{aligned} \tag{6}$$

Differentiating this expression yields

$$\frac{\partial L_i}{\partial \lambda_j} = \mu_i \frac{\partial E(x_i)}{\partial \lambda_j}. \tag{7}$$

Using (1) and the formula for expectation yields

$$\begin{aligned} \frac{\partial E(x_i)}{\partial \lambda_j} &= \sum_{x \in Z} \frac{\partial}{\partial \lambda_j} x_i \prod_{k=1}^{n} \frac{\rho_k^{x_k}}{x_k!} x(0) \\ &= \sum_{x \in Z} \left[\left(x_i x_j \frac{1}{\mu_j} \frac{\mu_j}{\lambda_j} \right) \pi(x) + x_i \frac{\pi(x)}{\pi(0)} \frac{\partial \pi(0)}{\partial \lambda_j} \right] \\ &= \frac{1}{\lambda_j} E(x_i x_j) + \frac{1}{\pi(0)} E(x_i) \frac{\partial \pi(0)}{\partial \lambda_j}. \end{aligned} \tag{8}$$

Using (2)

$$\begin{aligned} \frac{\partial \pi(0)}{\partial \lambda_j} &= -\pi^2(0) \frac{\partial}{\partial \lambda_j} \left[\sum_{x \in Z} \prod_{k=1}^{n} \frac{\rho_k^{x_k}}{x_k!} \right] \\ &= -\pi^2(0) \sum_{x \in Z} x_j \frac{1}{\mu_j} \frac{\mu_j}{\lambda_j} \frac{\pi(x)}{\pi(0)} \\ &= -\frac{\pi(0)}{\lambda_j} E(x_j). \end{aligned} \tag{9}$$

Substituting (9) into (8)

$$\frac{\partial E(x_i)}{\partial \lambda_j} = \frac{1}{\lambda_j} (E(x_i, x_j) - E(x_i)E(x_j)). \tag{10}$$

Finally, substituting (10) into (7) yields

$$\frac{\partial L_i}{\partial \lambda_j} = \frac{\mu_i}{\lambda_j}\left(E(x_i x_j) - E(x_i)E(x_j)\right).$$

Or

$$\frac{\partial L_i}{\partial \lambda_j} = \begin{cases} \frac{\mu_i}{\lambda_j}\ \mathrm{cov}(x_i, x_j), & \text{if } i \neq j \\ \frac{\mu_i}{\lambda_i}\ \mathrm{var}(x_i), & \text{if } i = j \end{cases}.$$

♦

B. Proof of Theorem 2

To help prove Theorem 2, we first prove three lemmas.

Lemma 2.1: The positive quadrant of the λ space, $\mathcal{R}_+^n$, maps into a region in the L space such that the outer boundary B of Y represents the infinite curve $\lambda = \infty$.[3]

Proof: Choose any λ. Consider an infinitesimal change $d\lambda$ from λ. From (3), this produces a change in L of:

$$\begin{bmatrix} dL_1 \\ \vdots \\ dL_n \end{bmatrix} = \begin{bmatrix} \frac{\mu_1}{\lambda_1}\ \mathrm{var}(x_1) & \cdots & \frac{\mu_1}{\lambda_n}\ \mathrm{cov}(x_1, x_n) \\ \vdots & & \vdots \\ \frac{\mu_1}{\lambda_n}\ \mathrm{cov}(x_1, x_n) & \cdots & \frac{\mu_n}{\lambda_n}\ \mathrm{var}(x_n) \end{bmatrix} \cdot \begin{bmatrix} d\lambda_1 \\ \vdots \\ d\lambda_n \end{bmatrix}.$$

The determinant of the matrix above $\neq 0$ unless:

1) the x_i are linearly dependent, namely $\exists(\alpha_1, \cdots, \alpha_n) \neq (0, \cdots, 0) \ni \alpha \cdot x = 0$ w.p. 1, or
2) $\mu_i \equiv \infty$ for some i, in which case the ith row contains all zeros, or
3) $\lambda_i \equiv \infty$ for some i, in which case the ith column contains all zeros.

Coordinate convexity of Z implies that Z must be a n-dimensional space (excepting the degenerate case $x_i \equiv 0$ for some i). Thus 1) can only be true if the state remains in some lower dimensional subset of the state space w.p. 1. This can only happen if the state remains in some portion of the boundary of Z, namely, if 2) or 3) holds. We ignore the degenerate case 2). Therefore, $\lambda \neq \infty$ implies that the determinant of the matrix above $\neq 0$. Thus, for any $\lambda \neq \infty$, we can choose a desired dL, and solve for the corresponding $d\lambda$ that produces it. Therefore, $\lambda \neq \infty$ corresponds to an L that must be in the interior of Y. Therefore, the boundary B of Y corresponds to $\lambda = \infty$.

Lemma 2.2: If $\lambda = \infty$, all states with nonzero probability lie on a hyperplane tangent to the state space Z (from above).

Proof: Any $\lambda = \infty$ can be written in the form

$$\lambda = \lim_{t \to \infty} \alpha t^{\gamma}$$
$$\text{for some } \gamma = (\gamma_1, \cdots, \gamma_n) \quad \text{where } \gamma_i \geq 0\ \forall\ i;$$
$$\text{and some } \alpha = (\alpha_1, \cdots, \alpha_n) \quad \text{where } \alpha_i \geq 0\ \forall\ i;$$
(where the multiplication is componentwise).

For any two states x' and x'', we have, from (1)[4]

$$\frac{\pi(x')}{\pi(x'')} \propto \frac{\Pi\lambda_i^{x'_i}}{\Pi\lambda_i^{x''_i}} = \Pi\lambda_i^{x'_i - x''_i}.$$

Using the limit form for $\lambda = \infty$ above,

$$\begin{aligned} \frac{\pi(x')}{\pi(x'')} &\propto \lim_{t \to \infty} \Pi t^{\gamma_i(x'_i - x''_i)} \\ &= \lim_{t \to \infty} t^{\Sigma\gamma_i(x'_i - x''_i)} \\ &= \lim_{t \to \infty} t^{\gamma \cdot x' - \gamma \cdot x''}. \end{aligned}$$

So if $\lambda = \infty$ and $\pi(x') \neq 0$, then

1) $\frac{\pi(x')}{\pi(x'')} > 0\ \forall\ x'' \in Z$ and thus $\gamma \cdot x' \geq \gamma \cdot x''\ \forall\ x'' \in Z$ and
2) if $\pi(x'') \neq 0$, then $\frac{\pi(x')}{\pi(x'')} < \infty$ and thus $\gamma \cdot x' = \gamma \cdot x''$.

Therefore, any state with nonzero probability lies on the hyperplane $\gamma \cdot x = \gamma \cdot x'$, which is tangent to the state space Z, from above, at x'.

Lemma 2.3: If all states with nonzero probability lie on a hyperplane, tangent to the state space Z at state x, and with normal vector γ, then L must lie on the hyperplane in the L space that passes through point $(\mu_1 x_1, \cdots, \mu_n x_n)$ with normal vector $(\mu_1\gamma_1, \cdots, \mu_n\gamma_n)$.

Proof: Equation (4) states that

$$L_i = \mu_i E(x_i) = \sum_{x \in Z} \mu_i x_i \pi(x).$$

The result follows from linearity.

Theorem 2: The feasibility region of L, Y, is a convex polyhedron.

Proof: By Lemma 2.1, the boundary B of Y corresponds to $\lambda = \infty$. By Lemma 2.2, this implies that all states with nonzero probability lie on a single hyperplane tangent to the state space Z. By Lemma 2.3, this corresponds to a point L on a corresponding hyperplane in the L space.

Therefore, the locus of points on the boundary B of Y is the union of hyperplanes corresponding to the hyperplanes tangent to the state space Z.

Now the hyperplanes tangent to the state space Z form the boundary of a convex polyhedron. Thus, B forms the boundary of a convex polyhedron. Thus, Y is a convex polyhedron.

Acknowledgment

The authors gratefully acknowledge helpful conversations with L. Ludwig, J. Walrand, and C. Dai.

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] J. S. Kaufman, "Blocking in a shared resource environment," *IEEE Trans. Commun.*, vol. COM-29, pp. 1474–1481, Oct. 1981.
[3] G. J. Foschini and B. Gopinath, "Sharing memory optimally," *IEEE Trans. Commun.*, vol. COM-31, pp. 352–360, Mar. 1983.
[4] J. T. Virtamo, "Reciprocity of blocking probabilities in multiservice loss systems," *IEEE Trans. Commun.*, vol. 36, pp. 1174–1175, Oct. 1988.

[3] By $\lambda = \infty$, we mean that $\lambda_i = \infty$ for some i.

[4] In the following, $\propto$ stands for "is proportional to," $\cdot$ represents a dot product, and all products and sums are taken over $i = 1, \cdots, n$.

[5] 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.

[6] F.P. Kelly, "Networks of queues with customers of different types," *J. Appl. Prob.,* vol. 12, pp. 542–554, 1975.

[7] F.P. Kelly, "Networks of queues," *Advances Appl. Prob.,* vol. 8, pp. 416–432, 1976.

[8] S.S. Lam, "Queueing networks with population size constraints," *IBM J. Res. Development,* pp. 370–378, July 1977.

[9] K.W. Ross and D.D. Yao, "Monotonicity properties for the stochastic knapsack," preprint, 1988.

[10] P. Nain, "Qualitative properties of the Erlang blocking model with heterogeneous user requirements," preprint, 1989.

[11] S. Ross, *Stochastic Processes.* New York: Wiley, 1983.

[12] F.P. Kelly, "Blocking probabilities in large circuit-switched networks," *Advances Appl. Prob.,* vol. 18, pp. 473–505, 1986.

[13] ——, "Routing in circuit-switched networks: Optimization, shadow prices, and decentralization," *Advances Appl. Prob.,* vol. 20, pp. 112–144, 1988.

[14] 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.

[15] B. Kraimeche and M. Schwartz, "Circuit access control strategies in integrated digital networks," in *Proc. IEEE Conf. Inform. Syst.,* 1984.

[16] S. Zachary, "Control of stochastic loss networks, with applications," *J. Royal Statis. Soc. B,* vol. 50, no. 1, pp. 61–73, 1988.

[17] W. Whitt, "Blocking when service is required from several facilities simultaneously," *AT&T Tech. J.,* vol. 64, pp. 1807–1856, 1985.

[18] 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.

[19] P. Bloom and P. Miller, "Intelligent Network/2," *Telecommun. Int.,* Feb. 1987.

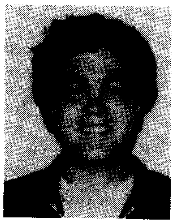
[20] Digest of Intelligent Networks Workshop, Lake Yamanaka, Japan, Oct. 1989.

[21] L. Ludwig, "A threaded/flow approach to reconfigurable distributed systems and service primitive architectures," *ACM SigCom,* Stow, VT, Aug. 1987.

[22] L. Pate, "IMAL analytical study: The interconnection of switches and concentrators with shared resources," *Bell Commun. Res.,* preprint, Mar. 1989.

[23] 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.

[24] F.P. Kelly, *Reversibility and Stochastic Networks.* New York: Wiley, 1979.



Scott Jordan received the B.S./A.B., the M.S., and PhD. degrees from the University of California, Berkely, in 1985, 1987, and 1990, respectively.

He is currently an Assistant Professor at Northwestern University. His teaching and research interests are the modeling and analysis of behavior, control, and pricing in computer/telecommunication networks, production, queueing, and other stochastic systems.

Pravin P. Varaiya (M'68–SM'78–F'80) received the B.S. degree from V.J.T. Institute, Bombay, India, and the M.S. and Ph.D. degrees from the University of California, Berkeley, all in electrical engineering.

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