Statistical Multiplexing and Mix-Dependent Alternative Routing in Multiservice VP Networks

Ching-Fong Su, *Member, IEEE,* and Gustavo de Veciana, *Member, IEEE*

Abstract—We consider problems in traffic integration and routing for virtual path (VP)-based multiservice networks. The objective is to exploit statistical multiplexing among various traffic types in order to improve system utilization. Difficulties arise due to statistical multiplexing since a connection's bandwidth requirement depends on the characteristics of the interfering traffic. We first consider whether segregating heterogeneous traffic with different quality of service (QoS) requirements on separate VP's is desirable. Next we consider routing heterogeneous permanent connections given a predefined traffic type mix onto multiple VP's between a source destination pair. We show that it is not necessarily advantageous to have each VP carry every traffic type. In fact, perhaps surprisingly, an optimum solution to this problem suggests that only a small number of traffic types, or even homogeneous traffic, need be present on each VP. Based on this observation, we propose a simple alternative routing algorithm with routing sequences depending on the traffic mix.

Index Terms—Alternative routing, statistical multiplexing, traffic mix, virtual path.

I. INTRODUCTION

ASYNCHRONOUS transfer mode (ATM) has been designed to meet the possible needs of integrated broadband communication networks. This technology is based on multiplexing and switching cells transported on virtual channel connections (VCC's). In turn, virtual path connections (VPC's) allow for joint handling of bundled VCC's and can serve as an effective way of reducing complex signaling and management tasks in a core network. The VP layer is in fact likely to serve as an intermediate resource management layer, wherein key resource allocation decisions are made on a somewhat slower time scale than typical connection times. Indeed, one can use the VP layer to simplify call admission control and routing and to segregate traffic based on QoS, traffic characteristics, and/or service classes.

VP capacity allocation is an important issue in managing VP-based networks. Indeed, if physical link capacities are divided into chunks (VP's), the utilization may be adversely affected by the resource segregation.[1] Hence the bandwidth allocated to each VP may need to adapt to the changes in traffic loads in order to improve the bandwidth utilization (see, e.g., [18] and [19]). In addition, the traffic mix on each VP also affects the bandwidth utilization because it affects the "efficiency" of statistical multiplexing. This paper first addresses the question of whether segregating heterogeneous traffic with different QoS requirements on separate VP's is desirable. The general belief is that one should segregate traffic with different QoS requirements on separate VP's (see e.g., [9]). However, this type of segregation is not necessarily "optimal." We shall see that traffic heterogeneity plays a critical role in multiplexing, and careful allocation of traffic mixes is essential to achieving good performance.

Similar care is needed in making routing decisions in a heterogeneous VP environment. There is much research and experience with routing policies in circuit-switched networks (see e.g., [1] and [7]), but one might ask if these principles will extend to multiservice networks. Routing schemes selecting the "least loaded path," e.g., real-time network routing (RTNR) of AT&T [1], can be useful in single service networks because they tend to balance the traffic load across the network and minimize the blocking probability (see, e.g., [8] and [12]). However, this may not be the case in a multiservice network. For example, it has been suggested that a "most loaded path" strategy might be preferable in order to leave room on the least-loaded path for future connections with high bandwidth requirements [22]. Various other routing policies, such as trunk reservation, have been investigated, and these are likely to also play a role in integrated services networks (see, e.g., [14]).

A simple example can illustrate the impact of statistical multiplexing and show how it might in turn affect routing policies. Consider a bufferless link shared by two types of traffic whose cell arrival rates are, for simplicity, modeled by Gaussian random variables with means m_1, m_2 and variances σ_1^2, σ_2^2, respectively. Suppose the link currently has n_1 and n_2 ongoing connections of each type. For this model, the bandwidth requirement, given a constraint on link overflow probability, is roughly [20], [21]

$$c(n_1, n_2) = (n_1 m_1 + n_2 m_2) + k\sqrt{(n_1 \sigma_1^2 + n_2 \sigma_2^2)}$$

C.-F. Su is with Fujitsu Laboratories of America, Inc., Sunnyvale, CA 94086 USA (e-mail: csu@fla.fujitsu.com).

G. de Veciana is with the Electrical and Computer Engineering Department, University of Texas at Austin, Austin, TX 78712 USA (e-mail: gustavo@ece.utexas.edu).

[1]In practice, buffer/bandwidth may be shared across VP's; however, herein we assume no such sharing.
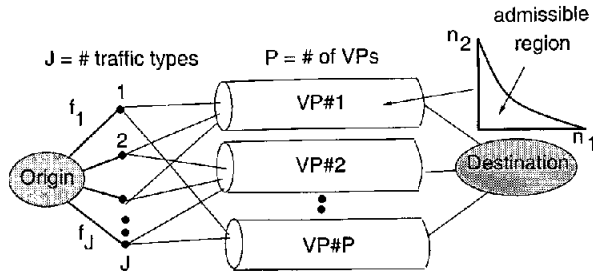
Fig. 1. Problem setup.

where k is a QoS parameter related to a link overflow probability. The bandwidth required for an additional connection of Type 1 can be approximated by

$$\frac{\partial c}{\partial n_1} \approx m_1 + \frac{1}{2}k\sigma_1^2\left(n_1\sigma_1^2 + n_2\sigma_2^2\right)^{-\frac{1}{2}}. \qquad (1)$$

Note that the marginal bandwidth needed for an additional connection depends on the traffic characteristics, mix, and load. For example, given the mean m and the variance σ^2 of a connection's cell arrival rate, the key factor determining the marginal bandwidth requirement is the variance of the aggregate traffic $n_1\sigma_1^2 + n_2\sigma_2^2$ currently on the links it traverses. Based on this example, one could argue that it may be "cheaper" (consuming less additional bandwidth) to route a new connection through a link whose current aggregate variance is large. Moreover, by selecting routes with minimum marginal bandwidth requirements, one could make more resources available to incoming connections or other types of services. Alternatively, from a system perspective, the network routing mechanisms might "tailor" the traffic mix across the network resources in order to achieve better overall efficiency. These two arguments are in sharp contrast to typical routing policies that try to balance the loads on the network.

In this paper, we consider the problem of routing multiple traffic types onto multiple VP's between an origin-destination pair (see Fig. 1). Such variety in VP's over distinct physical links can be used to increase reliability or may be the result of capacity constraints. In our problem formulation we shall consider two types of constraints: those on the numbers of connections that can be admitted to each VP and one characterizing the mix of traffic to be carried. In general, the relationship among the admissible numbers of connections of each type on a VP is nonlinear resulting from the characteristics of statistical multiplexing. Our constraint on the traffic mix is intended to capture the desired relative traffic loads or operation point of the network. Such constraints might be obtained from optimizations accounting for the demand and revenue generated by each traffic type.

Alternative routing algorithms will be considered for the network shown in Fig. 1. In general, *alternative routing* comes in many flavors depending on how the routing sequence is selected (see, e.g., [7]). We shall consider routing sequences that account for statistical multiplexing and traffic mix on the VP's. In short, different routing sequences are assigned to different traffic types so as to maintain a desirable traffic mix on each VP and to enhance statistical multiplexing.

The balance of this paper is organized as follows. In Section II we discuss the role of statistical multiplexing and nonlinear boundaries of call admission regions associated with VP's carrying heterogeneous traffic. The problem of whether traffic with different QoS requirements should be segregated on separate VP's is analyzed in Section III. Routing issues are discussed in Section IV and are followed by simulations and conclusions.

II. ADMISSIBLE REGION OF A VP

In general, to assess the exact admissible numbers of connections on a given link, one would have to resort to either simulation or a significant amount of computation. In this paper, we will resort to a common (conservative) approximation based on Chernoff's bound in the context of bufferless multiplexing (see, e.g., [6], [10], and [15]). Similar approximation can also be obtained for buffered systems (see, e.g., [4]). Most of the ideas in this paper follow from the *qualitative* characteristics of the admissible region resulting from this bound.

Suppose N i.i.d. traffic streams are carried on a bufferless link, and each stream has a stationary cell arrival rate $A_i, i \in \{1, \ldots, N\}$. Assume that the link capacity is c and we require that the aggregate arrival rate to the link exceeds the capacity only rarely—with a probability no larger than δ. Based on Chernoff's bound [2], the overflow probability is upper bounded by

$$P\left(\sum_{i=1}^{N} A_i > c\right) \leq \exp\left(-\sup_{\theta > 0}[c\theta - N\Lambda(\theta)]\right) \qquad (2)$$

where $\Lambda(\theta) = \log(E[\exp(\theta A_i)])$ is the logarithm of the moment generating function of A_i.

One can generalize the result to analyze statistical multiplexing of heterogeneous traffic. Let

$$A = \sum_{j=1}^{J} \sum_{i=1}^{n_j} A_{ji}$$

be the aggregate arrivals, where A_{ji} are independent random variables denoting the cell arrival rate of the i-th traffic flow of Type j, and $\Lambda_j(\theta) = \log(E[\exp(\theta A_{ji})])$. In this case, Chernoff's bound and the overflow constraint give the following requirement:

$$\log P(A > c) \leq -\sup_{\theta > 0}\left[c\theta - \sum_{j=1}^{J} n_j\Lambda_j(\theta)\right] \leq \log\delta. \qquad (3)$$

Following [15] we shall define the *admissible region* $\mathcal{A}(c)$ for a bufferless link with capacity c as the collection of vectors $\mathbf{n} = (n_1, \ldots, n_J) \geq 0$ that satisfy (3), i.e.,

$$\mathcal{A}(c) = \left\{\mathbf{n} \mid \mathbf{n} \geq 0 \text{ and } \sum_{j=1}^{J} \alpha_j(\theta)n_j - \frac{\log(\delta)}{\theta} \leq c \right. \\ \left. \text{for some } \theta > 0\right\} \qquad (4)$$

where $\alpha_j(\theta) = \Lambda_j(\theta)/\theta$.

We note, as argued in [11], that the complement of the admissible region in R_+^J is a convex set. Thus a linear approximation
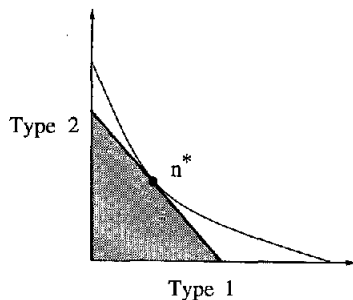
Fig. 2. Linearization of the admissible region's boundary.

could be used to conservatively represent the boundary of the admissible region (see e.g., [10] and [15]). A linearization of $\mathcal{A}(c)$ can be done by fixing an operating point $\mathbf{n}^*$ on the boundary [2] and using a tangent hyperplane at $\mathbf{n}^*$ as a linear approximation to the boundary (see Fig. 2). Associated with each $\mathbf{n}^*$, there exists a θ^* that achieves the sup in (3), so that we can express the linearized region by

$$\mathcal{A}_{\mathbf{n}^*}(c) = \left\{ \mathbf{n} \mid \mathbf{n} \geq 0 \text{ and } \sum_{j=1}^{J} \alpha_j n_j \leq c' \right\} \quad (5)$$

where $\alpha_j = \alpha_j(\theta^*)$ is the "effective bandwidth" of Type j traffic and $c' = c + (\log(\delta)/\theta^*)$ is the link's "effective capacity" with respect to θ^* (i.e., a linearization at $\mathbf{n}^*$). Given such a linearization, the relative bandwidth requirements of each traffic type become fixed, resulting in a linearly constrained region and making call admission straightforward. However, a key remaining question is where to linearize the admissible region boundary.

Given the nonlinearity of the admissible region's boundary, it may be important to select the operating point $\mathbf{n}^*$ carefully.[3] As shown in (5), the effective bandwidth α_j of each traffic type depends on $\mathbf{n}^*$. Thus in principle, the network can exploit this dependence to change the relative resource requirements of the various traffic types in order to improve the overall resource utilization. We will show in Section IV that a careful selection of such operating points can result in a better statistical multiplexing and reduce call blocking probability.

III. INTEGRATION OR SEGREGATION OF TRAFFIC?

It is generally believed that one should segregate traffic with different QoS requirements on their own VP's (see, e.g., [9]). The intuition is that if we multiplex traffic with different QoS requirements on the same VP, then the overall QoS for the VP should be the most stringent QoS requirement. By providing a QoS that is more stringent than necessary to some traffic streams, we waste network resources. However, due to the nature of statistical multiplexing, it may still be more "economical" to put all traffic on the same VP. In the following, we use an example with two traffic types to illustrate the roles

[2]In this paper, we refer to the possibly nonlinear constraints (3) of the admission region as the boundary.

[3]Note, however, that the admissible region's boundary becomes more "linear" when the link capacity increases (see, e.g., [23]).

of statistical multiplexing and traffic mix when the QoS is related to cell losses resulting from link overflows.

Suppose there are N total connections, which consist of a fraction f_1 of Type 1 flows and a fraction f_2 of Type 2 flows, where $f_1 + f_2 = 1$.[4] In order to get a qualitative understanding, we shall first resort to Gaussian traffic models for which an explicit expression for bandwidth requirements exists and then consider the popular On/Off traffic model. The cell arrival rates of each traffic type are modeled by Gaussian random variables with mean and variance (m_1, σ_1^2), (m_2, σ_2^2), respectively. We assume that the two traffic types are carried by a bufferless link and require cell loss ratios of 10^{-6} and 10^{-3}, respectively. The problem is to decide whether to partition a link into two segregated VP's, or form a single shared VP, where the goal is to minimize the required bandwidth.

The bandwidth requirements of a single shared VP and segregated VP's are shown in (6) and (7), respectively

$$c_1 = N(f_1 m_1 + f_2 m_2) + k_1 \sqrt{N} \sqrt{f_1 \sigma_1^2 + f_2 \sigma_2^2} \quad (6)$$

$$c_2 = N(f_1 m_1 + f_2 m_2) + \sqrt{N} \left[k_1 \sqrt{f_1 \sigma_1^2} + k_2 \sqrt{f_2 \sigma_2^2} \right] \quad (7)$$

where k_1 and k_2 are the QoS parameters. Indeed, for the Gaussian model, the tail distribution of aggregate cell rates (i.e., the overflow probability) can be captured by the deviations from mean, and k_1, k_2 are multiples of the standard deviation [20], [21]. Without loss of generality, we assume $k_1 > k_2$. For the aforementioned QoS of 10^{-6} and 10^{-3}, we have that k_1 and k_2 are 4.7534 and 3.0902, respectively. We are interested in a condition making $c_1 \leq c_2$, so it is advantageous to form a single VP and give Type 2 traffic a better QoS rather than setting up two VP's. In other words, the benefit of statistical multiplexing outweighs the loss in overprovisioning for a better QoS. Surprisingly, the condition depends only on σ_1/σ_2 and f_1, where N plays no role. Based on (6) and (7), we can show that $c_1 \leq c_2$ if

$$\frac{\sigma_1}{\sigma_2} \geq \frac{k_1^2 - k_2^2}{2k_1 k_2} \sqrt{\frac{1 - f_1}{f_1}}. \quad (8)$$

Hence to decide whether to integrate two types of traffic on the same VP, one needs to assess if the ratio of their standard deviations exceeds a threshold, which depends on their QoS requirements and the traffic mix.

In Fig. 3, we plot the threshold on σ_1/σ_2, i.e., the right side of (8), as a function of the traffic mix f_1 with the aforementioned k_1 and k_2. The threshold defines the *integration* and *segregation* regions. For example, for $\sigma_1/\sigma_2 = 0.6$, we should form a single VP when the fraction of Type 1 traffic exceeds 0.35. Otherwise, it is more efficient to have two VP's with different QoS.

Based on the regions shown in Fig. 3, it is clear that when f_1 is small, the ratio of σ_1/σ_2 needs to be large in order to make integration beneficial. An interpretation for this might be that we waste a larger amount of bandwidth in bringing a better QoS to Type 2 traffic (i.e., integration) when f_1 is small. Thus only

[4]For simplicity we assume f_1, f_2 are real numbers even though they should be restricted to multiples of $1/N$ such that Nf_1, Nf_2 are integers.
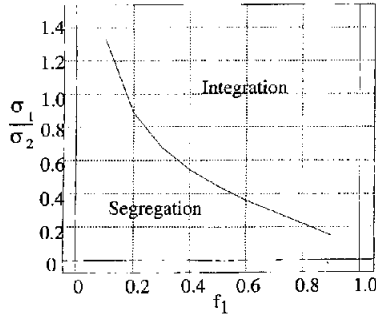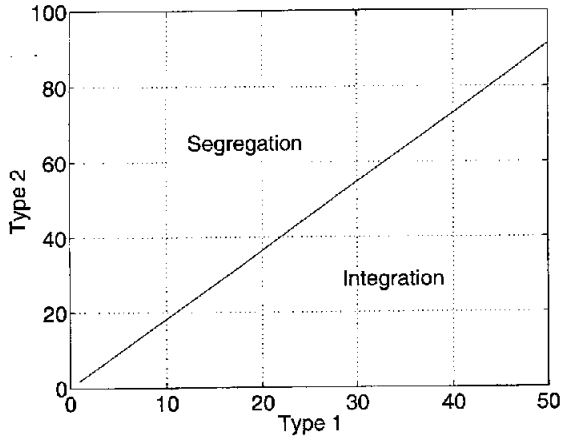
Fig. 3. Traffic mix versus σ_1/σ_2.

Fig. 4. Integration and segregation regions.

Fig. 5. Integration and segregation regions.

when Type 1 traffic is "bursty," i.e., has high variance, can the benefit of better multiplexing outweigh the waste of bandwidth. Therefore, the threshold on σ_1/σ_2 should be larger when f_1 is small, as shown in Fig. 3. By contrast, when f_1 is large, Type 2 traffic is less significant, so the threshold becomes less stringent, i.e., *integration* is desirable. These tradeoffs are captured by the curve in Fig. 3.

Note that the term $(k_1^2 - k_2^2)/2k_1k_2$ in (8) serves as a scaling factor for the curve in Fig. 3. If k_1 and k_2 are close, then the threshold on σ_1/σ_2 for integration becomes small, which increases the *integration* region. That is, one is indeed likely to integrate traffic with similar QoS requirements when minimizing the overall bandwidth requirements.

One can consider the problem of whether to integrate or segregate from a different perspective. Suppose n_1 and n_2 are the numbers of Type 1 and Type 2 connections carried by the system. Inequality (8) can then be written as

$$\frac{n_1\sigma_1^2}{n_2\sigma_2^2} \geq \left[\frac{k_1^2 - k_2^2}{2k_1k_2}\right]^2.$$

In Fig. 4, we plot the *integration* and *segregation* regions with respect to n_1 and n_2 with the aforementioned k_1, k_2, and $\sigma_1/\sigma_2 = 0.6$. The figure clearly indicates that the ratio of n_1 and n_2, i.e., the mix, rather than their magnitudes, determines the decision boundary.

Note that although the total number of connections N does not play a role in the condition (8) determining whether it is beneficial to form a single VP, the bandwidth savings $\Delta c = c_2 - c_1$ grows as $\sqrt{N}$ as the number of connections increases. Let us
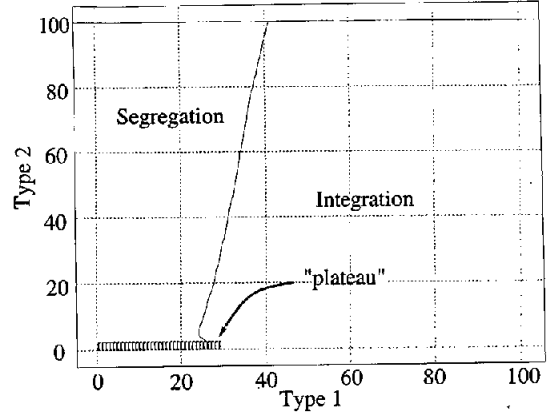
call Δc divided by $N(f_1m_1 + f_2m_2)$ the *normalized* bandwidth savings resulting from integration. It is clear that the *normalized* bandwidth savings are proportional to $1/\sqrt{N}$, i.e., they become less significant as N increases. Indeed, when N is large, the *effective bandwidth* of each traffic stream approaches its mean rate for both the integrated and segregated VP's. Hence the (normalized) difference between (6) and (7) becomes smaller as N increases.

Next we consider On/Off traffic models. For such models, it is difficult to derive a simple criterion such as (8) to determine whether to segregate or integrate heterogeneous traffic. The complexity is mainly due to the lack of a closed-form expression for the bandwidth requirement. Nevertheless, by using (3), we can exhibit such tradeoffs.

Suppose we have two types of On/Off traffic with mean rates 0.5, 0.15 and peak rates 0.9, 0.8, respectively. In addition, the overflow probability requirements are respectively 10^{-9} and 10^{-3}. As shown in Fig. 5, there exist *integration* and *segregation* regions, which depend on the traffic mix (n_1, n_2). As in the Gaussian case, the traffic mix determines whether it is advantageous to integrate the two traffic types. An interesting observation is the small "plateau" region around $(n_1 = 0, n_2 = 0)$ in Fig. 5. This arises because the number of connections is too small to achieve any statistical multiplexing given the specified QoS. In fact, in this region, bandwidth reservation would be done based on peak-rate allocation. This phenomenon was pointed out in [5]. The author argued that in order for the traffic to be "statistically multiplexable," the total allocated bandwidth for a single type of aggregated traffic needs to exceed a critical value. Similarly, a minimum number of connections is required to see the "economies of scale." Hence it makes no difference whether the two traffic types are integrated or segregated when the number of connections is small. Note that for Gaussian traffic, there is no such threshold in the number of connections.

As with Gaussian traffic, the desirability of traffic integration for On/Off traffic depends on the traffic mix and traffic characteristics, but now also the load. These results show that accounting for the efficiency of multiplexing is essential in order to minimize the total bandwidth reservation. For cases with more than two traffic types, these observations still hold, but a simple criterion for decision making is unlikely due to the

increased complexity. Nevertheless, optimal VP configurations for multiple traffic types could be determined numerically or by simulation.

IV. ROUTING IN MULTISERVICE VP NETWORKS

Next we consider routing connections on the simplified network shown in Fig. 1. We assume that the goal is either to maximize the number of admissible permanent connections or to minimize the call blocking probabilities in a dynamic system, subject to a load with a predefined traffic mix and the capacity constraints on each VP. We also assume that all VP's provide the same aggregate QoS, which are equal to or better than those requested by each traffic type. As shown in Section III, integrating traffic and providing a better QoS on VP's could be advantageous.

A. Routing Permanent Connections

The problem of determining the maximum admissible number of permanent connections can be formulated as a linear programming problem where we relax the integer constraint on the number of connections. If the total number of connections of each type are big enough, the rounding errors to the closest integers will be negligible. Thus an optimum solution to the linear programming problem is close to an optimum solution of the *integer* programming problem.

Suppose there are J traffic types to be routed onto P virtual paths with linearized admissible regions, where $\alpha_{p,j}$ is the effective bandwidth of Type j traffic on VP p and all VP's can carry all traffic types.[5]

Let $n_{p,j}$ denote the number of Type j connections carried by VP p. The capacity constraints are then given by the following inequalities:

$$\sum_{j=1}^{J} \alpha_{p,j} n_{p,j} \leq c_p \quad \forall p \tag{9}$$

where c_p is the effective capacity of VP p (see (5)).

Next we consider a set of constraints associated with the predefined traffic mix. Let $f_j, j = 1, \ldots, J$ be the fraction of each traffic type and $\sum_{p=1}^{P} n_{p,j}$ denote the total number of Type j connections carried by all VP's. To ensure the traffic mix constraint is met, we require that

$$\frac{\sum_{p=1}^{P} n_{p,1}}{f_1} = \frac{\sum_{p=1}^{P} n_{p,2}}{f_2} = \cdots = \frac{\sum_{p=1}^{P} n_{p,J}}{f_J}. \tag{10}$$

We can rewrite these as a set of equality constraints[6]

$$\sum_{p=1}^{P} n_{p,1} - \frac{f_j}{f_1} \sum_{p=1}^{P} n_{p,j} = 0 \quad \forall j \neq 1. \tag{11}$$

---

[5]For simplicity the dependence of $\alpha_{p,j}$ on θ^* has been omitted. Nevertheless, it should be clear that $\alpha_{p,j}$ depends on the selected linearization of admissible region.

[6]We assume that $\sum_{p=1}^{P} n_{p,j}$ are big enough such that the error of rounding $n_{p,j}$ to the closest integers are negligible, so the traffic mix can be roughly maintained.

# traffic types (J)	# VP (P)	# homogeneous VP
2	2	≥ 1
2	3	≥ 2
2	4	≥ 3
3	3	≥ 1
3	4	≥ 2

The goal is to find an allocation that maximizes the total number of connections $\sum_{j=1}^{J} \sum_{p=1}^{P} n_{p,j}$ subject to the constraints (9) and (11). Given the nature of this optimization, the constraints in (9) will hold with equality since we have relaxed $n_{p,j}$ to be continuous. If equality did not hold on VP p, then $n_{p,j}, j = 1, \ldots, J$ could be proportionally increased such that traffic mix constraints are not violated. This linear programming problem can then be formulated as follows:

$$\max \quad \sum_{j=1}^{J} \sum_{p=1}^{P} n_{p,j}$$

$$\text{subject to} \quad \sum_{j=1}^{J} \alpha_{p,j} n_{p,j} = c_p \qquad \forall p$$

$$\sum_{p=1}^{P} n_{p,1} - \frac{f_j}{f_1} \sum_{p=1}^{P} n_{p,j} = 0 \quad \forall j \neq 1 \tag{12}$$

$$n_{p,j} \geq 0 \qquad \forall p, j.$$

The constraints can be represented in a compact form as $A\vec{n} = b$, $\vec{n} \geq 0$, where $\vec{n} = \{n_{p,j} \mid p = 1, \ldots, P, j = 1, \ldots J\}^T$ is a vector representing the number of connections on every VP. In addition, we let $\mathbf{n}_p = \{n_{p,j} \mid j = 1, \ldots, J\}^T$ denote the number of connections on VP p.

Theorem 1: Suppose $P, J > 1$; then there exists a solution to the permanent route assignment problem (12) in which at least $P - \min[P, J - 1]$ VP's carry homogeneous traffic.

Proof:: There exists a solution, say, $\vec{n}^*$, which is a vertex of the polytope corresponding to the feasible set [13], [17]. In addition, it can be shown that a vertex of the feasible set contains at most rank(A) nonzero components [13], [17]. Note that A is an $M \times N$ matrix, where $M = P + J - 1$, $N = PJ$, and rank(A) $\leq M$. Thus, $\vec{n}^*$ contains at least $N - M$ zero components, where a zero component, e.g., $n_{p,j} = 0$, means that VP p does not carry Type j traffic.

Based on the number of nonzero components in $\vec{n}^*$, we can find a lower bound on the number of VP's carrying homogeneous traffic. Since each VP will carry some traffic, there exists at least one nonzero component of $\vec{n}^*$ for each VP. Therefore, at most rank(A) $- P \leq J - 1$ additional positive $n_{p,j}^*$ can be spread out across the P VP's. Thus the number of homogeneous VP's is greater than $P - \min[P, J - 1]$. $\square$

If $J - 1 < P$, then some VP's could carry homogeneous traffic. Table I shows this "lower bound" on the number of homogeneous VP's for various numbers of VP's and traffic types. For instance, when two traffic types are routed onto three VP's, we find that at least two VP's could carry homogeneous traffic.

This observation sheds light on how the traffic could be carried by the VP's in order to maximize the admissible number of

connections subject to a predefined mix. In particular, the theorem implies that one can allocate routes to connections such that the following are true:

- fraction of heterogeneous VP's $\leq \min[1, (\# \text{Types-}1/\# \text{VP's})]$;
- average number of traffic types per VP $\leq 1 + (\# \text{Types-}1/\# \text{VP's})$.

Thus as the ratio of numbers of traffic types to VP's $(\# \text{Types}/\# \text{VP}) = J/P$ decreases, more virtual paths could carry homogeneous traffic.

The key insight is that it is not necessarily advantageous for each VP to carry *all* traffic types in this heterogeneous setup. To maximize the throughput, only a small number of traffic types or even homogeneous traffic will need be present on each VP. This suggests that, in practice, optimized multiservice networks with sufficient routing diversity might end up looking like multiple logical networks, which are "segregated" by service type. As we will see in Section IV-B, strict segregation is not advantageous when connections arrive and leave dynamically; however, approximately maintaining this regime will improve performance.

1) Nonlinear Admissible Region: Theorem 1 was shown for the case where the VP's had linear admissible regions; however, one can show that the same result holds if the admissible region for each VP has a convex boundary (see, e.g., Fig. 2). This is discussed in Appendix A.

B. Dynamic Routing

Suppose that calls of each traffic type arrive as Poisson processes with rates λ_j and each type of call has an arbitrary holding time distribution with mean μ_j^{-1}. We will consider the routing problem with the objective of minimizing blocking probabilities. The results in Section IV-A suggest that connections need to be allocated to VP's carefully in that appropriate traffic mixes need to be maintained on each VP in order to maximize the total admissible number of connections. However, it is not always possible to maintain desirable traffic mixes on each VP in a dynamic environment. In the following, we propose: 1) a simple multiservice alternative routing scheme, which approximately maintains the traffic mix on each VP around the desirable operating points associated with the permanent connection routing assignment and 2) a complementary algorithm to compute the routing sequences to be attempted for each traffic type based on the offered traffic mix.

1) *Multiservice alternative routing:* Each traffic type is assigned a *sequence* of VP choices. A new connection of a given type is routed to the first VP on the sequence associated with its type, which can carry the connection.

2) *Mix-dependent routing sequences:* Given the offered loads for each traffic type $\rho_j = \lambda_j/\mu_j$, let $f_j = \rho_j/(\sum_j \rho_j)$ denote the fraction of Type j traffic. One can then solve the permanent route assignment problem in Section IV-A to find a desirable operating point $\bar{n}^*$ on a vertex of the (possibly linearized) feasible set. Such solutions will tend to maximize the number of homogeneous VP's. The sequence of VP's to be at-

tempted for Type j traffic is then based on $\bar{n}^*$ according to the following rules:

- VP p precedes VP q on the sequence if $\bar{n}^*$ is such that VP p carries only Type j traffic and VP q carries heterogeneous traffic.
- VP p precedes VP q on the sequence if either (1) $n_{p,j}^* > n_{q,j}^*$, or (2) $n_{p,j}^* = n_{q,j}^*$ and $p < q$ (break ties).

For example, consider a three VP network carrying three traffic types where the components of $\bar{n}^*$ are given by $(n_{1,1}^*, n_{1,2}^*, n_{1,3}^*) = (7, 28, 0)$, $(n_{2,1}^*, n_{2,2}^*, n_{2,3}^*) = (28, 0, 0)$, and $(n_{3,1}^*, n_{3,2}^*, n_{3,3}^*) = (0, 7, 36)$. The sequence of VP's to be used by the routing algorithm are $2 \rightarrow 1 \rightarrow 3$, $1 \rightarrow 3 \rightarrow 2$, and $3 \rightarrow 1 \rightarrow 2$ for Types 1, 2, and 3 respectively. In particular, a Type 1 connection would first attempt routing on VP 2, then VP 1, and last VP 3.

1) A Repacking Routing Policy: In general, the order in which calls arrive will affect the routing decisions and thus the efficiency of the network. Theoretically, a routing policy might recompute the best routes whenever the system status changes, e.g., new connections arrive. In particular, such a policy might reroute (repack) connections in order to admit a new request. By doing so, the network admits connections in a greedy fashion without regard to connection types or network revenue. Although such a policy is not necessarily "optimal," it does attempt to aggressively admit connections and thus reduce overall blocking probabilities. We will use such a policy as a baseline for comparison in evaluating various routing algorithms.

V. SIMULATION RESULTS

In this section, we use simulations to evaluate the performance of the proposed algorithm under two scenarios: 1) two traffic types on two VP's and 2) three traffic types on three VP's. We compare the call blocking probabilities achieved by our mix-dependent alternative routing to a variety of routing algorithms including least loaded route (LLR) and minimum-resource route (MRR), as well as the repacking policy.

A. Two Traffic Types and Two VP's

We first consider the case of two traffic types on two VP's. As shown in Table I, we found that if we were routing permanent connections, at least one VP could carry homogeneous traffic. This property leads to two possible routing VP sequences, i.e., a sequence $1 \rightarrow 2$ for one type and $2 \rightarrow 1$ for the other type. As a result, each type is assigned a distinct primary VP. Connections first attempted on the primary VP for that type, and then the alternative is tried. A connection is blocked if the second trial fails.

The proposed algorithm is compared with other algorithms under various traffic loads, which are denoted by ρ_1 and ρ_2 respectively. In our simulations, both traffic types are assumed to be On/Off with peak rates 1 and 0.5 as well as mean-to-peak ratios 10% and 90%, respectively. The two VP's have an identical
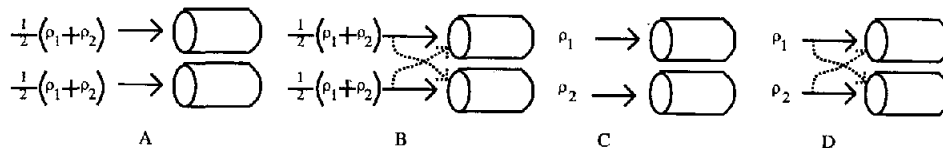
Fig. 6. Four routing algorithms.

TABLE II
THE COMPARISON OF BLOCKING PROBABILITIES. NOTE: N/A DENOTES THE BLOCKING PROBABILITIES THAT ARE SMALLER THAN 10^{-5}

Loads (ρ_1, ρ_2)		A	B	C	D	E	LLR	MRR
(40,40)	B_1	0.0597	0.03099	0.00101	0.00221	0.00081	0.03024	0.02891
	B_2	0.0837	0.05027	0.01869	0.00222	0.00158	0.05149	0.04283
(45,30)	B_1	0.0269	0.00791	0.00729	0.00044	0.00013	0.00760	0.00720
	B_2	0.0420	0.01477	0.00022	0.00048	0.00014	0.01580	0.01212
(30,45)	B_1	0.0498	0.02280	N/A	0.00148	0.00076	0.02258	0.02257
	B_2	0.0627	0.03149	0.05410	0.00149	0.00137	0.03201	0.02767
(50,25)	B_1	0.0197	0.00462	0.02657	0.00069	0.00033	0.00438	0.00419
	B_2	0.0326	0.00913	N/A	0.00095	0.00037	0.00981	0.00765
(25,50)	B_1	0.0542	0.02796	N/A	0.00289	0.00169	0.02867	0.02971
	B_2	0.0655	0.03405	0.10479	0.00290	0.00279	0.03458	0.02990

capacity of 25 and provide an overflow probability of 10⁻⁹. The admissible region of each VP is the nonlinear region resulting from the Chernoff's bound (see (4)). The routing algorithms we compare are shown in Fig. 6 and explained below.

- *Balanced-load scheme without retrial:* Connections of each type are sent to each VP with equal probability. The loads sent to each VP are $(1/2)(\rho_1 + \rho_2)$.
- *Balanced-load scheme:* This is the same as Algorithm A except that if the selected VP is unavailable, the other VP is tried. That is, each connection is assigned a routing sequence $1 \to 2$ or $2 \to 1$ with equal probability regardless of its type.
- *Segregated-load scheme:* Connections of each type are sent to distinct VP's. If the selected VP is unavailable, the connection is blocked.
- *Proposed algorithm:* Different traffic types are assigned different routing sequences: $1 \to 2$ or $2 \to 1$.
- *Repacking scheme:* Connections are repacked if this will permit a new connection to be admitted.

We also simulated two additional dynamic routing algorithms: LLR and MRR. In LLR, a newly arriving connection is sent to the VP that has the largest free capacity.[7] By contrast, in MRR, a connection is sent to the VP where minimum additional bandwidth is required to carry the new connection given the current load and statistical multiplexing. If the selected VP is unavailable, the other VP is tried.

1) Comparison of Blocking Probabilities: The simulation results are shown in Table II where B_1 and B_2 are the blocking probabilities of Type 1 and Type 2 traffic, respectively. Table II shows that Algorithm A has worse overall blocking probabilities than Algorithm B because it uses static load sharing without retrying the other path if the selected path is unavailable. Hence a connection may be blocked unnecessarily. Similarly, Algorithm C did not utilize the VP capacity efficiently either, i.e.,

each VP carries only one traffic type. We found that a VP might experience significant call blocking while the other VP has unused capacity.

Excluding the repacking algorithm (E) and the unacceptable case of Algorithm C, Algorithm D has the smallest overall blocking probabilities for both traffic types under all traffic load combinations. This confirms our observations made in the context of routing permanent connections, namely, that unbalanced traffic mixes on the VP's will improve the efficiency of statistical multiplexing and the overall usage of VP capacities. Notice that the call blocking probabilities of Algorithm D are close to those achieved by the repacking algorithm (E). The traffic load ratios on each VP of Algorithm B are roughly equal to the original offered loads ratio ρ_1/ρ_2, so multiplexing is not as efficient as that in Algorithm D. For example, when the traffic load is (45, 30), the blocking probabilities of Algorithm D are 0.00044 and 0.00048, which are more than a order of magnitude smaller than those achieved by Algorithm B.

Our simulations show that LLR and MRR are also inferior to the proposed algorithm, and their blocking probabilities are close to those of Algorithm B for both types of traffic. Since LLR and MRR fail to account for the traffic mix in making routing decisions, the traffic load ratios are roughly equal to the original offered load ratio ρ_1/ρ_2 on each VP. Therefore, LLR and MRR cannot achieve as efficient a multiplexing as the proposed algorithm does, which leads to higher blocking probabilities. These observations provide insight into how one should extend the routing policies in circuit-switched networks to multiservice packet-switched networks. Obviously, the effect of the traffic mix and multiplexing on the selected routes should be accounted for when routing connections.

B. Three Traffic Types and Three VP's

Next we compare two alternative routing algorithms with the repacking scheme for a network supporting three traffic types on three VP's. For simplicity, we assume the admissible regions of

---

[7]The free capacity is the difference between VP's capacity and the minimum bandwidth required to carry the connections already on the VP.

TABLE III
CAPACITY AND THE EFFECTIVE BANDWIDTHS

VP	Capacity	Effective Bandwidth		
		Type 1	Type 2	Type 3
1	100	2	2.5	3
2	100	3	4	5
5	100	2	2	2

TABLE IV
COMPARISON OF BLOCKING PROBABILITIES

Traffic loads (ρ_1, ρ_2, ρ_3)		E	F	G
(33,33,33)	B_1	0.00150	0.00868	0.00174
	B_2	0.00190	0.01180	0.00238
	B_3	0.00190	0.01418	0.00306
(42,42,15)	B_1	0.00080	0.00461	0.00129
	B_2	0.00110	0.00613	0.00179
	B_3	0.00110	0.00748	0.00223
(60,24,24)	B_1	0.00340	0.01773	0.00528
	B_2	0.00430	0.02329	0.00724
	B_3	0.00430	0.02825	0.00891
(50,30,15)	B_1	0.00035	0.00321	0.00055
	B_2	0.00045	0.00429	0.00077
	B_3	0.00035	0.00531	0.00094
(90, 9, 9)	B_1	0.00240	0.00637	0.00226
	B_2	0.00240	0.00818	0.00301
	B_3	0.00240	0.00983	0.00370
(3,45,45)	B_1	0.00270	0.00692	0.00266
	B_2	0.00370	0.01027	0.00425
	B_3	0.00440	0.01265	0.00547

VP's are linearized. The "effective bandwidth" and the "effective capacity" of the VP's are shown in Table III. The following algorithms were considered.

- *Repacking scheme:* Connections are repacked if this will permit a new connection to be admitted.
- *Balanced-load scheme:* Three sequences of possible VP choices, $1 \rightarrow 2 \rightarrow 3$, $2 \rightarrow 3 \rightarrow 1$, and $3 \rightarrow 1 \rightarrow 2$, are randomly assigned to each connection with equal probability regardless of the traffic type. Hence the VP's are attempted in a rotary fashion.
- *The proposed algorithm:* Each connection is assigned a sequence of possible VP choices based on its type. The sequences are computed using the rules discussed in Section IV-B.

The blocking probabilities achieved by these algorithms are shown in Table IV. As expected, the balanced-load algorithm (F) resulted in the worst overall performance. For example, when the traffic loads are (50, 30, 15), then the blocking probabilities for Type 1 traffic, B_1, are 0.00035 and 0.00055 for Algorithms E and F, respectively, while Algorithm G's, 0.00321, was almost an order of magnitude worse. This discrepancy in blocking probabilities was consistent across a variety of loads. In addition, the blocking probabilities of the proposed algorithm (G) are pretty close to those of Algorithm E. These results show that mix-dependent routing algorithms can significantly decrease blocking probabilities.

C. Robustness, Linearization, and "Resource Pooling"

In our proposed algorithm, the routing sequences of each traffic type depend on the *relative* magnitudes of the compo-

TABLE V
ROUTING SEQUENCE UNDER VARIOUS TRAFFIC LOADS

Traffic loads (ρ_1, ρ_2, ρ_3)	Routing Sequence		
	Type 1	Type 2	Type 3
(15,30,60)	$2 \rightarrow 1 \rightarrow 3$	$1 \rightarrow 2 \rightarrow 3$	$3 \rightarrow 1 \rightarrow 2$
(18,30,60)	$2 \rightarrow 1 \rightarrow 3$	$1 \rightarrow 2 \rightarrow 3$	$3 \rightarrow 1 \rightarrow 2$
(15,30,72)	$2 \rightarrow 1 \rightarrow 3$	$1 \rightarrow 2 \rightarrow 3$	$3 \rightarrow 1 \rightarrow 2$
(15,30,48)	$2 \rightarrow 1 \rightarrow 3$	$1 \rightarrow 2 \rightarrow 3$	$3 \rightarrow 1 \rightarrow 2$
(18,24,60)	$2 \rightarrow 1 \rightarrow 3$	$1 \rightarrow 2 \rightarrow 3$	$3 \rightarrow 1 \rightarrow 2$
(12,24,72)	$2 \rightarrow 1 \rightarrow 3$	$2 \rightarrow 1 \rightarrow 3$	$3 \rightarrow 1 \rightarrow 2$

nents of $\vec{n}^*$. When the offered traffic loads (mix) are perturbed, the sequences are usually preserved. For example, when the traffic loads in Section V-B are perturbed around an operational point (15, 30, 60) by 20%, most of the computed routing sequences stay unchanged (see Table V). In fact, only the routing sequence of Type 2 traffic changes when traffic loads are perturbed from (15, 30, 60) to (12, 24, 72). Hence the proposed routing algorithm is likely to be robust to small fluctuations of traffic load or measurement errors.

Suppose the admissible regions for all VP's were linearized in the same fashion, that is, the effective bandwidth for each traffic type is identical on each VP. In this case, resource requirements for a given connection are neither VP nor mix dependent so routing questions reduce to determining a VP that can carry the connection. Performance differences among algorithms that attempt all available VP's in any order will be insignificant up to the possible effects of integrality constraints. In this scenario, we have *resource pooling* (see, e.g., [16]), in the sense that the various VP's between a given source destination pair perform like a big VP with the sum of their individual effective capacities. In practice, one may then argue that it might be advantageous to linearize the admissible regions identically on all resources in order to simplify routing. However, this simplicity will come at the cost of reduced network utilization, particularly for relatively small VP's, or shared resources, which are expected to carry radically different traffic mixes.

In [3], an argument was made that the boundaries of admissible regions can be linearized without significantly reducing the achievable "revenue" in high-capacity networks. Nevertheless, unless the linearization points are chosen carefully and appropriate routing policies are used to keep the networks around the desired operational regime, linearization can lead to a loss in overall revenue. Note that the desirable linearization point $\vec{n}^*$ depends on the offered loads for each traffic types as well as the capacities of the various VP's.

VI. CONCLUSION

In this paper, we have considered problems related to resource allocation and routing in integrated services networks. In particular, we were motivated by questions that arise in managing heterogeneous traffic types with possibly different QoS requirements using VPC's as an intermediate resource management layer.

The first question that arose was whether heterogeneous traffic with different QoS requirements should be segregated on distinct VP's or aggregated on a single VP but given the most stringent QoS requirement. Based on a simple model,

our analysis showed that the answer is certainly not straightforward, i.e., in some cases, it is advantageous to aggregate while in others it is better to segregate. A criterion for making such decisions was derived for Gaussian models, exhibiting the dependence on the traffic characteristics, traffic mix, and QoS requirements. Similar behavior is likely to hold for more general setups, where the essential tradeoff is between achieving improved statistical multiplexing by aggregating versus losing efficiency due to provisioning for the most stringent QoS.

Next we considered the problem of routing multiple traffic types with a predefined traffic mix onto multiple VP's between a source destination pair. Our analysis and simulations suggested that it is critical to account for heterogeneous statistical multiplexing or traffic mixes in making routing decisions. In fact, to maximize the throughput, only a small number of traffic types, or even homogeneous traffic are present on each VP. This suggests that, in practice, multiservice networks with sufficient routing diversity might end up operating like multiple logical networks, which are segregated by service type.

Based on our observations, we proposed an alternative routing scheme with mix-dependent routing sequences, which aims to maintain appropriate traffic mixes on each VP in order to increase the overall system utilization. In addition to the routing scheme, we also propose an algorithm to compute the routing sequence depending on traffic mix and VP capacity. The proposed routing scheme results in a significantly improved performance over LLR and is robust to variations in the offered traffic mix. Last, we make the following key point: by linearizing the admissible set for network resources in the same fashion, one can trade off network utilization with routing "complexity" in multiservice networks.

APPENDIX
EXTENDING THEOREM 1 TO NONLINEAR ADMISSIBLE REGIONS

Suppose admissible regions for each VP have convex boundaries. In order to determine an optimal routing of permanent connections, we can formulate an optimization problem as in (12) where the capacity constraints are now determined by the nonlinear boundaries associated with each VP's admissible set $\mathcal{A}_p$. Notice that the feasible set F associated with these capacity constraints is no longer a polytope. Nevertheless, there still exists an optimal allocation $\vec{n}^*$, which maximizes the objective function $f(\vec{n}) = \sum_{j=1}^{J} \sum_{p=1}^{P} n_{p,j}$ and satisfies the capacity as well as the traffic mix constraints.

Note that $\vec{n}^*$ is necessarily located at the boundary of feasible set F. Hence the components of $\vec{n}^*$ consist of P points (i.e., n_p^*) and each n_p^* is on the boundary of $\mathcal{A}_p$. With respect to the points n_p^*, one can linearize $\mathcal{A}_p$ to obtain an approximated region $\mathcal{A}_p'$ (see e.g., Fig. 2). Based on these linearized admissible regions, we can formulate a linear programming problem similar to (12) with a new feasible set F'. Notice that these linearized admissible regions $\mathcal{A}_p'$ are contained inside the original admissible regions $\mathcal{A}_p$ (except for the points n_p^*) due to the convexity assumption of the boundaries (see, e.g., Fig. 2). Thus it is clear that $F' \subset F$.

Next we consider two cases: either $\vec{n}^*$ is a vertex of F' or $\vec{n}^*$ is not a vertex of F'. In the latter case, there exists a vertex $\vec{v}$ of F' such that $f(\vec{v}) \geq f(\vec{n}^*)$ [13], [17]. However, $f(\vec{v}) \leq f(\vec{n}^*)$ because $\vec{n}^*$ is a maximizer of $f(\cdot)$ over F and $\vec{v} \in F' \subset F$. Thus we have that $f(\vec{v}) = f(\vec{n}^*)$. Therefore, we can always find a vertex of F' which is also a maximizer of $f(\cdot)$ w.r.t. F. It then follows, as in Theorem 1, that the such a maximizer would have at least $P - \min[P, J - 1]$ homogeneous VP's.

REFERENCES

[1] G. R. Ash, "Dynamic network evolution, with examples from AT&T's evolving dynamic network," *IEEE Commun. Mag.*, vol. 33, pp. 26–39, 1995.
[2] P. Billingsley, *Probability and Measure*. New York: Wiley, 1986.
[3] S. Borst and D. Mitra, "Asymptotically achievable performance in ATM networks," *Adv. Appl. Prob.*, to be published.
[4] N. G. Duffield and N. O'Connell, "Large deviations and overflow probabilities for the general single-server queue, with applications," *Math. Proc. Cam. Phil. Soc.*, vol. 118, pp. 363–374, 1995.
[5] A. Elwalid, D. Mitra, and R. H. Wentworth, "A new approach for allocating buffers and bandwidth to heterogeneous, regulated traffic in an ATM node," *IEEE J. Select. Areas Commun.*, vol. 13, no. 6, pp. 1115–1127, 1995.
[6] R. J. Gibbens and F. P. Kelly, "Measurement-based connection admission control," presented at the 15th Int. Teletraffic Congr., 1997.
[7] A. Girard, *Routing and Dimensioning in Circuit-Switched Networks*. Reading, MA: Addison-Wesley, 1990.
[8] S. Gupta, K. Ross, and M. E. Zarki, "Routing in virtual path based ATM networks," in *Proc. IEEE Globecom*, 1992, pp. 571–575.
[9] S. Gupta, K. W. Ross, and M. El Zarki, *Routing in Communications Networks*, M. Steenstrup, Ed. Englewood Cliffs, NJ: Prentice-Hall, 1995, ch. 2.
[10] J. Y. Hui, "Resource allocation for broadband networks," *IEEE J. Select. Areas Commun.*, vol. 6, pp. 1598–1608, 1988.
[11] J. Y. Hui, *Switching and Traffic Theory for Integrated Broadband Networks*. Boston, MA: Kluwer Academic, 1990.
[12] R.-H. Hwang, "LLR routing in homogeneous VP-based ATM networks," in *Proc. IEEE Infocom 95*, 1995, pp. 587–593.
[13] H. Karloff, *Linear Programming*. Boston, MA: Birkhauser, 1991.
[14] F. P. Kelly, "Routing and capacity allocation in networks with trunk reservation," *Math. Oper. Res.*, vol. 15, no. 4, 1990.
[15] ——, "Effective bandwidths of multi-class queues," *Queueing Syst.*, vol. 9, no. 1, pp. 5–16, 1991.
[16] F. P. Kelly and C. N. Laws, "Dynamic routing in open queueing networks: Brownian models, cut constraints and resource pooling," *Queueing Syst.*, vol. 13, pp. 47–86, 1993.
[17] D. G. Luenberger, *Linear and Nonlinear Programming*. Menlo Park, CA: Addison-Wesley, 1984.
[18] S. Ohta and K. Sato, "Dynamic bandwidth control of the virtual path in an asynchronous transfer mode network," *IEEE Trans. Commun.*, vol. 40, pp. 1239–1247, 1992.
[19] A. Orda, G. Pacifici, and D. E. Pendarakis, "An adaptive virtual path allocation policy for broadband networks," in *Proc. IEEE Infocom'96*, vol. 1, 1996, pp. 329–335.
[20] A. Papoulis, *Probability & Statistics*. Englewood Cliffs, NJ: Prentice-Hall, 1990.
[21] I. Sidhu and S. Jordan, "Multiplexing gains in bit stream multiplexors," *IEEE/ACM Trans. Networking*, vol. 3, pp. 785–797, 1995.
[22] R. Siebenhaar, "Multiservice call blocking approximations for virtual path based ATM networks with CBR and VBR traffic," in *Proc. IEEE Infocom 95*, 1995, pp. 321–329.
[23] C.-F. Su, "Efficient Traffic Management Based on Deterministically Constrained Traffic Flows," Ph.D. dissertation, Electrical and Computer Engineering Dept., Univ. of Texas at Austin, 1998.
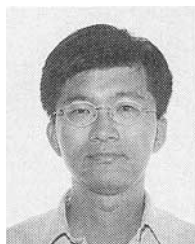
Ching-Fong Su (M'98) received the B.S. degree in electrical engineering from National Tsing-Hua University, Taiwan, R.O.C., in 1991 and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Texas at Austin in 1995 and 1998, respectively.

Since 1998, he has been a Member of Research Staff with Fujitsu Labs of America Inc. His current research interest includes routing in high-speed networks, multimedia applications, and performance evaluation of communication systems.
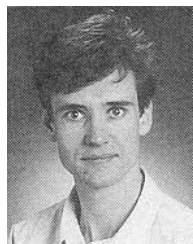
Gustavo de Veciana (S'88–M'94) received the B.S., M.S., and Ph.D. degrees in electrical engineering from the University of California at Berkeley in 1987, 1990, and 1993, respectively.

In 1993, he joined the Department of Electrical and Computer Engineering at the University of Texas at Austin, where he is currently an Associate Professor. His research focuses on issues in the design and control of telecommunication networks.

Dr. de Veciana is an Editor for the IEEE/ACM TRANSACTIONS ON NETWORKING. He received a General Motors Foundation Centennial Fellowship in electrical engineering and a 1996 National Science Foundation CAREER Award.