ATM망에서의 멀티미디어 트래픽 제어

안 병준, 이 형호

bjahn@etri.re.kr, holee@etri.re.kr

한국전자통신연구원 라우터기술연구부

Multimedia traffic management in ATM networks

Byungjun Ahn, Hyeong Ho Lee

Router Technology Department, ETRI

Abstract

The problem of bandwidth allocation and routing in VP based ATM networks was studied. A priori reservation of resources for VP's reduces the statistical multiplexing gain, resulting in increased Call Blocking Probability (CBP). The focus of this study is on how to reduce CBP by the efficient bandwidth allocation and routing algorithms. Equivalent capacity concept was used to calculate the required bandwidth by the call. And the effect of traffic dispersion was explored to achieve more statistical gain. A cost-effective traffic dispersion routing algorithm, CED, was designed. The algorithm finds the optimal number of dispersion paths for a call, where the gain balances the dispersion cost. Simulation study showed that CED could significantly reduce the CBP.

1. Introduction

Recently we have seen so called "traffic explosion" generated by millions of customers who are using new Internet services (e.g., World Wide Web, etc.). More bursty and bandwidth-hungry new services are expected to emerge in near future. Beside the demand for the tremendous amounts of network capacity for high-quality transmissions, in their nature, traffic classes generated by these services tend to have heterogeneous traffic characteristics and different Quality of Service (QoS) requirements (e.g., cell loss rate, delay, and jitter, etc.). Thus efficient management of network resources (e.g., bandwidth allocation, etc.) becomes a rigorously difficult task for network engineers.

Asynchronous Transfer Mode (ATM) is the transfer mode for implementing Broadband Integrated Services Digital Networks (B-ISDNs) and other high-speed networks. ATM provides a unified interface that is based on 53 octet cells for a variety of services having harshly different requirements. ATM cells are routed through fixed paths. Links and nodes in the network are shared by means of bandwidth allocation. Bandwidth allocation deals with the problem of determining the amount of bandwidth required by a connection for the network to provide the required QoS. In general, two different bandwidth allocation schemes are used in ATM: deterministic and statistical multiplexing. For bursty traffic sources, statistical multiplexing is desirable to achieve high utilization of network resources. Most statistical bandwidth allocation schemes are based on the well-known concept of effective bandwidth that has been studied extensively. Among many schemes, equivalent capacity (or effective bandwidth used in this study), based on the work by Guérin *et al.* [1], provides bandwidth requirement of a single or multiplexed connections on the basis of their statistical characteristics. In other words, for a given a QoS requirement (i.e., cell loss probability in this study) and a few traffic descriptors for each traffic source, equivalent capacity represents the minimum bandwidth needed at the multiplexer to support an arbitrary collection of traffic sources together without violating the QoS requirements.

Traffic dispersion is credited as an effective way to improve link utilization and network performance, especially when peak-to-mean ratio and peak-to-link capacity ratio of the burst are relatively high [2],[3]. By using traffic dispersion, a burst is divided into many sub-bursts that are transmitted in parallel through multiple paths and are resequenced at the destination. In this study, we refer to traffic dispersion as a cyclic spreading of cells from a source over available paths. Throughout the work done so far, there is no thorough report on efficient traffic dispersion algorithms and the impact of traffic dispersion on the effective capacity, when link capacity is reserved on VP's. These are issues on which we focus in this study. Equivalent capacity is used to calculate the required bandwidth by the call. Each call is represented as bursty and heterogeneous multimedia traffic. First, we explore the

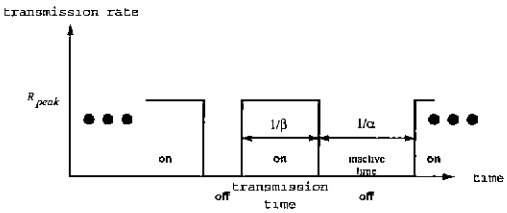
effect of the traffic dispersion to achieve more statistical gain. Through this study, it is discovered how the effect of traffic dispersion changes with different traffic characteristics and the number of paths. Based on what are discovered, a cost effective traffic dispersion algorithm is designed. Simulation study shows that the algorithm can significantly reduce CBP.

The organization of this paper is as follows. Next section presents how traffic dispersion affects the equivalent capacity. A cost-effective traffic dispersion algorithm is presented in section 3, while simulation design and results are discussed in section 4. Section 5 summarizes the findings of this work and outlines possible future work.

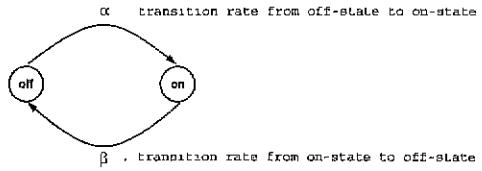
2. Effect of traffic dispersion on equivalent capacity

In ATM networks, Call Admission Control (CAC) handles bandwidth allocation. Estimating the required statistical bandwidth is of most important in any CAC strategy. The statistical bandwidth of a connection depends not only on its own stochastic characteristics but also strongly on the characteristics of existing connections that share the link capacity. There has been much research effort on the methods of evaluating the required bandwidth of aggregate ATM traffic sources. Among them, equivalent capacity, proposed by Guérin *et al.* [1], outperforms other schemes in many cases [4],[5]. In their method, the equivalent capacity of the aggregate ATM traffic sources is determined by the minimum of either fluid-flow approximation or stationary approximation. The fluid-flow approximation accurately estimates the equivalent capacity when the impact of individual connection's traffic characteristics is critical. It is a useful tool in many situations because of its additive property. Given any set of multiplexed sources, one calculates the equivalent capacity of each source, and simply adds the capacities. In fluid-flow approximation, the equivalent capacity of each traffic source is independent on the characteristics of other traffic sources. The assumption here is that each source requires the same QoS parameter (i.e., cell loss ratio in our study). However, when a large number of bursty connections are multiplexed together, their aggregated statistical behavior differs from their individual traffic characteristic. This leads the fluid-flow approximation to a conservative estimate of the equivalent capacity required. For such a case, stationary approximation provides reasonably accurate estimate, which approximates the distribution of the stationary bit rate on a link. As both approximations overestimate the actual value of the equivalent capacity for different range of connections characteristics, the equivalent capacity is taken to be the minimum of fluid-flow approximation and stationary approximation to predict the relatively accurate equivalent capacity of connections.

The concept of equivalent capacity can be generalized to a variety of traffic source models, including those such as the Markov-modulated models that are frequently used to describe voice and video signals. For the sake of simplicity, however, we assume that each traffic source is represented by the two-state, continuous-time Markov chain of Fig. 1.



(a) Two-state on-off traffic source model



(b) Markov fluid-flow model

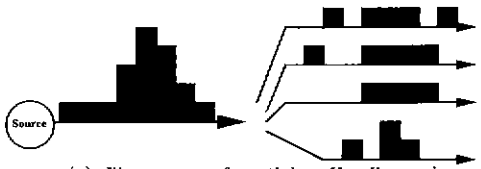
Fig. 1 Basic on-off source model

This model, with appropriate adjustment of parameters, can be used to describe voice, compressed (VBR) video, and image traffic, as well as other bursty traffic. It has been shown that the equivalent capacity expression, obtained using this simple two-state on-off source model, is a special case of the more general form of Markov fluid-flow model [6]. It should be also noted that the traffic sources don't all have to be the on-off type. The additivity of equivalent capacities applies to any set of traffic sources sharing the same delay and loss probability QoS parameters for which one can define effective capacities.

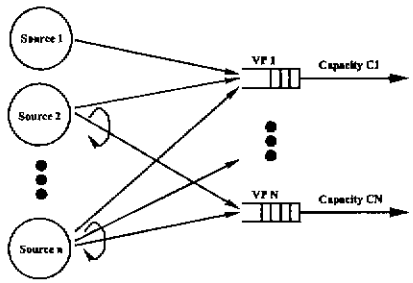
For a reasonably well-connected network there would be several paths from a sender to any given destination. It may be necessary, for instance, to provide reliability. The total capacity is partitioned spatially over the paths. Traditionally, only one of them would be chosen for the information transfer; usually the shortest one, measured in actual length or number of hops. Instead of using a single path, the sending process might disperse its traffic over all the paths leading to the desired destination. A resource sharing close to the optimal would then be possible. We call this generic technique "*traffic dispersion*". Fig. 1 shows an illustration of spatial traffic dispersion and a model of a one stage multiplexer where k sources generate traffic that is spread over N links.

Traffic dispersion makes it possible to alleviate the effects of bursty traffic and equalize the network load without introducing the delay incurred by shaping. The technique applies equally well to datagram as to VC networks. The traffic can be dispersed over multiple paths in the network, multiple links within a path, or multiple physical channels,

such as frequency or wavelength channels, within a link. The important thing in order to yield the gain is that the paths, links, or channels do not share transmission capacity statistically. Traffic is dispersed cell by cell or burst by burst over the chosen set of routes. The cells are put back in order at the receiver if needed. Traffic dispersion is akin to alternate path (multipath) routing, which provides several possible paths from which to choose other alternative path when the optimal path becomes congested for some reason. For example if a node notices congestion on its primary path, it reroutes the traffic over a precomputed alternate path. The distinction we make between traffic dispersion and alternate path routing is that the latter is done on the time scale of a session, while dispersion is done on cell or bursts of cells within a session. In most instances dispersion is preventive, while alternative path routing is reactive, triggered by some network problem.



(a) Illustration of spatial traffic dispersion



(b) One-stage multiplexer

Fig. 2 Generic traffic dispersion technique

Following assumptions were made for this study.

- Although the optimal number of VP's between a pair of nodes is a subject to design, any pair of nodes has one or more VP's for each different class of QoS requirement. Thus, each VP carries only one type of traffic with a specific QoS requirement.
- In this study, we do not explicitly address the topology design problem. Instead, we assume that any pair of nodes in the network has one or more VP's for each different class of QoS requirement. Since a VP network is very likely to be connected densely, we believe this assumption is feasible.
- We use deterministic VP capacity reservation strategy. In this study, we assume that the reallocation of VP capacity should be done periodically on a much longer time scale than the interarrival time of successive calls. Furthermore, we assume that the time interval between two VP capacity reallocations is significantly larger than the VC setup time. Under this assumption, routing of VC's can be performed as if the topology and the capacity of the VP network were fixed.
- We consider direct routing algorithms. In direct routing, calls are allowed to be routed only via a direct VP between a source node and a destination node.
- Heterogeneous traffic is considered. Each call behaves as an on-off fluid source represented by the two-state continuous-time Markov chain. Successive on-off periods (i.e., $1/\beta$: mean burst length, and $1/\alpha$: mean idle time) are assumed to be mutually independent and identically distributed. Alternatively, an on-off traffic source is represented by three parameters: connection's peak rate R_{peak}, utilization of the connection $\rho=\alpha/(\alpha+\beta)$, and mean of the burst period b.
- Call arrivals follow Poisson arrival and call durations are exponentially distributed.
- We define link load as the aggregate bandwidth of all VP's passing through the link. Throughout this study, equivalent capacity, proposed by Guerin *et al.* [1], is used to estimate bandwidth requirement.
- The traffic dispersion strategy used in this study is dividing source peak rate evenly by the number of VP's involved (i.e., cyclic dispersion).

In the following, we illustrate how traffic dispersion affects the equivalent capacity needed for the transmission of heterogeneous traffic. With dispersion, the traffic from each source is sent over a separate path, disjoint from all the other paths. Each path is therefore only affected by the traffic from one of the dispersed sources, and this source can be seen as the fraction of traffic that the original source sends over that specific path. We define the *dispersion factor*, N, as the number of paths over which the traffic from a source is spread. For an on-off traffic source with peak rate R_{peak} cells/unit-time, cyclic dispersion of cells corresponds to reducing the peak rate of a source on each of the paths to R_{peak}/N cells/unit-time, while source utilization ρ and mean burst period b are kept same.

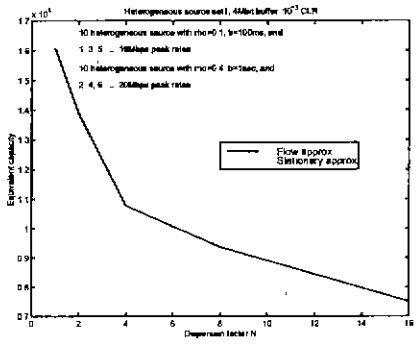


Fig. 3 Effect of traffic dispersion on equivalent capacity

Including the case shown in Fig. 3, a number of different cases were examined In general, equivalent capacity decreases as N grows. However, simulation results show that a dispersion factor of about eight seems to be sufficient. At that point, most of benefits are obtained, and increasing the dispersion factor more than eight does not give significant improvements. Furthermore, dispersion causes larger capacity reductions in the case with a relatively small or moderate buffer size. For very large buffers dispersion does not affect the equivalent capacity, but will probably reduce the delay, and for very small buffers dispersion over a modest number of paths cannot improve the situation, unless there are enough sources to obtain multiplexing effects Since traffic dispersion requires resequencing and extra signaling to setup multiple VC's, it should be used only when it gives significant benefits. This is the basic idea for our design of a cost effective dispersion algorithm. The algorithm finds an optimal value of N, where the gain in equivalent capacity (i.e., capacity reduction) balances the traffic dispersion cost.

3. Cost Effective Dispersion

The traffic dispersion algorithm proposed in this study, named Cost Effective Dispersion (CED), decides the optimal number of paths based only on the current statistics of the VP load and the traffic descriptor of new call. If someone can speculate the statistical characteristics of upcoming calls, more efficient algorithm could be designed. However, heterogeneous input traffic, as modeled in the previous section, does not lend us the long-term perspective of it. This is because:

- The equivalent capacity required by a VP varies very sensitively depending on the traffic descriptors of the calls already existing on that particular VP as well as the traffic descriptor of the new call at the instance of call arriving.
- Even without using traffic dispersion, statistics of VP load will vary significantly as a result of Call Admission Control (CAC) depending on the sequence of call arrivals.
- Traffic dispersion alters the characteristics of input traffic by spreading out the peak rate into multiple VP's
- Traffic dispersion algorithm is initiated upon arrival of a new call by the statistical multiplexer of a VP source node, and the decision must be made on the fly.

As a consequence, in the design process of the algorithm, major concern had to be on the optimization of the cost-performance function that decides the number of dispersion paths upon arriving of a new call. The algorithm determines whether traffic dispersion is used or not. When traffic dispersion is beneficial, it determine the favorable paths in terms of the cost of traffic dispersion. The proposed traffic dispersion algorithm, as shown in Fig. 4 is optimized for both the traffic dispersion cost and the call level QoS (i.e., CBP). Because traffic dispersion requires resequencing and multiple setup of VC's, it should be used selectively only when it is necessary to keep the overall CBP below the given call level QoS. Efficiency of the algorithm was evaluated in terms of the probability of dispersion and the average number of paths in this study. The algorithm is designed to increase the number of dispersion paths only when the reduction in the equivalent capacity (i.e., traffic dispersion gain) is greater than the cost measured in equivalent capacity.

The cost of the traffic dispersion is a linear function of the number of dispersion paths N, and defined as:

cost = average free capacity x ***coefficient*** x (N-1),

where ***coefficient*** is given as an input parameter of the algorithm.

Assume that VP_1, VP_2, ..., VP_i, ..., VP_M are available.
$\Delta=\infty$; $S=\{\varnothing\}$;
N = 1;

do

  for each VP_i **do**
    calculate $\delta_{N,i}$ induced by the new call with (R_{peak}/N, ρ, b);
  end
$S_N=\{\phi\}$;
select N VP's with the least $\delta_{N,i}$ -> S_N;
$\Delta_N=\sum\delta_{N,i}$
if $N \geq 2$ **then**
  if $\Delta_N < \Delta_J$ –cost and $\Delta_N < \Delta$ **then**
    $\Delta = \Delta_N$; $S=S_N$;
  endif
else
    $\Delta = \Delta_J$; $S=S_J$;
endif

$N=N+1$;

while $N \leq M$

Fig. 4 Cost Effective Dispersion Algorithm

By taking free capacity into account, traffic dispersion algorithm is more adaptive to the network load. In lightly loaded network, traffic dispersion is used only when the dispersion gain is significant while it is more likely used even with smaller gain in heavily loaded network. For the sake of simplicity of designing and managing a statistical multiplexer, it is assumed that the source peak rate is divided equally when traffic dispersion is used. Cells are generated by a source at its peak rate R_{peak}, and they are transmitted through N dispersion paths in cyclic manner so that each

dispersion path receives cells at the rate of R_{peak}/N. By dividing source peak rate R_{peak}, equivalent capacity requirement induced by the new call is effectively distributed to N dispersion paths. For each dispersion path, however, the increment in equivalent capacity by the source peak rate R_{peak}/N is quite different from that of other dispersion paths depending on the traffic descriptors of existing calls on that path. Traffic dispersion gain does not always increase monotonically nor linearly as the number of dispersion paths increases. It totally depends on the statistical distribution of traffic descriptors of existing calls on each VP, which varies at each VP. Simulation results show that the traffic dispersion gain is not considerable when the number of dispersion paths is larger than 8.

4. Simulation Results

4.1 Simulation parameters

For the purpose of comparison, two non-dispersion routing algorithms (i.e., LLP and mIS) are tested. Upon call arrival, LLP selects a least-loaded VP_i while mIS chooses a single path VP_i with the least δ_i. As a dispersion routing algorithm, CED is used. Following parameters were used for traffic descriptors.

- R_{peak}: exponentially distributed with mean values of 8Mbps and 16Mbps
- $0 < \rho \leq 1$: exponentially distributed with mean values of 0.1 and 0.2
- $b > 0$: exponentially distributed with mean values of 0.5 sec and 1.0 sec

4.2 Performance of routing algorithms

Typical simulation results is presented in Fig. 5, where average R_{peak}=8Mbps, ρ=0.1, b=0.5*sec*, and mean Poisson arrival rate λ=1.24 were used.

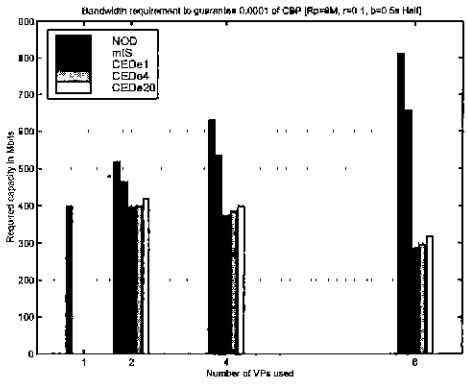


Fig. 5 Bandwidth requirement

Simulation results including this case show following characteristics. Effect of b is not significant. This observation is a strong indication that the required equivalent capacity is mostly determined by stationary approximation when the average number of connections is relatively large. Even though both LLP and mIS are single path algorithms, their performances are quite different. This is because the equivalent capacity varies sensitively depending on the statistical characteristics of existing calls. When either LLP or mIS is used, for all simulation results, required capacity increases as the number of VP's grows. This is coincident with the argument that *a priory* reservation of resources on VP's reduces the statistical multiplexing gain, resulting in an increased CBP

Traffic dispersion is particularly effective when multiple VP's are used. In those cases, simulation results show that CED can save 40%~60% of capacity. Even when the physical link capacity is large enough to establish a single VP with huge capacity, traffic dispersion can save about 30% of capacity in many cases.

4.3 Effect of CED coefficient on the dispersion factor

Fig. 6 illustrates the effect of CED cost coefficient on the dispersion factor D_f, the number of paths taken by a call. These were measured when simulations in Fig. 2 were performed. In this particular instance, bandwidth requirements are about same, no matter what coefficient value is used. However, as intended, the statistics of dispersion factor differs when different CED cost coefficients is used: smaller the coefficient, larger the dispersion factor. With the coefficient of 0.05, only 30% of calls took more than one paths when 8 VP's were used, 10% when 4 VP's were used. For other input traffic characteristics, bandwidth requirements differ up to 20%, depending on the coefficient. It does not increase linearly as a function of the coefficient: rather it varies with the number of VP's as well as the given input traffic characteristics. Although it is not linear, bandwidth requirement increases as the coefficient increases.

Thus, network engineers can have an option to choose from less dispersion and less bandwidth requirement.

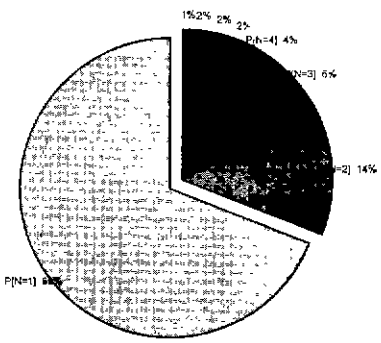


Fig. 6 Effect of CED coefficient on the dispersion factor

4.4 Traffic characteristics as a result of routing

When input traffic is routed to multiple VP's or dispersed, statistical characteristics of the input traffic seen by each VP are much different from those of input traffic that initially arrived at the system (i.e., the multiplexer). Thus, in the following, we investigate the effect of routing algorithm on the characteristics of traffic. In particular, we are interested in the distributions of interarrival time, peak rate, source utilization and mean burst period. These are very important variables when we develop analytical models. Thus, we explore these quantities more intensively later.

CDF of interarrival time seen by each VP is exponentially distributed and is analogous to that of input traffic initially arrived at the system. Note that, before routing, interarrival time of input traffic is exponentially distributed with mean of 1/1.24. Mean and standard deviation of exponential distributions should be same. When CED is used, these two are about same. From Fig. 3 the mean number of paths taken by a call is found to be 3.86. If we assume that each VP is selected equally likely, the mean call interarrival time at each VP is 1.67, same as the one measured by simulation. Thus, the assumption is proven to be correct. Similar results were obtained when 4 VP's are used.

Same arguments are possible for the distribution of peak rates. When either mIS or LLP is used, mean of peak rate distribution is almost equal to that of input traffic before routing. For CED, mean of peak rate is divided by 3.86, the average number of paths taken by a call. However, variance was reduced, meaning that the distribution is not exponential one. We will investigate this in detail later. For mIS and LLP, R_{peak} distribution at individual VP was same as that of input traffic.

The distributions of source utilization ρ and mean burst period b were kept unchanged, as expected. When CED is used, means of these two distributions are slightly increased. When mIS is used, mean of mean burst period distribution differs from VP to VP, in particular, for the case of 4 VP's.

5. Conclusion

The problems of bandwidth allocation and routing in Virtual Path (VP) based Asynchronous Transfer Mode (ATM) networks were studied. As an efficient way to facilitate the network management, VP concept has been proposed in the literature. Traffic control and resource management are simplified in VP based networks. However, a priori reservation of resources for VP's also reduces the statistical multiplexing gain, resulting in increased Call Blocking Probability (CBP). The focus of this study was on how to reduce CBP (or equivalently, how to improve the bandwidth utilization for a given CBP requirement) by the effective bandwidth allocation and routing algorithms.

Equivalent capacity concept was used to calculate the required bandwidth by the call. Each call was represented as bursty and heterogeneous multimedia traffic.

First, the effect of traffic dispersion was explored to achieve more statistical gain. No other work in the literature did thorough investigation of traffic dispersion algorithm capable of finding the optimal number of dispersion paths depending on the dynamic link load, when heterogeneous multimedia traffic is applied. This was an issue on which we focused in this study. Through this study, it was discovered how the effect of traffic dispersion varies with different traffic characteristics and the number of paths. Efficient routing algorithm CED was designed. Since traffic dispersion requires resequencing and extra signaling to set up multiple VC's, it should be used only when it gives significant benefits. This was the basic idea in our design of CED. The algorithm finds an optimal dispersion factor for a call, where the gain balances the dispersion cost. Simulation study showed that CED could significantly reduce the CBP when network resources are allocated to multiple VP's. As intended, the statistics of dispersion factor differs when different CED cost coefficient is used. smaller the coefficient, larger the dispersion factor. It was also shown that the bandwidth required to guarantee the given QoS, does not increase linearly as a function of the coefficient, rather it varies with the number of VP's as well as the given input traffic characteristics. Although it is not linear, bandwidth requirement increases as the coefficient increases. Thus, network engineers can have an option to choose from less dispersion and less bandwidth requirement. Next, this study provided analysis of the statistical behavior of the traffic seen by individual VP, as a result of traffic dispersion. This analysis is essential in estimating the required capacity of a VP accurately when both multimedia traffic and traffic dispersion are taken into account.

References

[1] R. Guérin, H. Ahmadi, and M. Naghshineh, "Equivalent Capacity and Its Application to Bandwidth Allocation in High-Speed Networks," *IEEE Journal on Selected Areas in Communications*, vol. 9, no. 7, pp. 968-981, Sept. 1991.

[2] E. Gustafsson and G. Karlsson, "When Is Traffic Dispersion Useful? A Study On Equivalent Capacity," in *ATM Networks: Performance Modeling and Analysis* (D.D. Kouvatsos, ed.), vol. 2, pp. 110-129, New York, New York: Chapman & Hall, 1996.

[3] E. Gustafsson and G. Karlsson, "A Literature Survey on Traffic Dispersion," *IEEE Network*, vol. 11, no. 2, pp. 28-36, March/April 1997.

[4] H. G. Perros and K. M. Elsayed, "Call Admission Control Schemes: A Review," *IEEE Communications Magazine*, vol. 34, no. 1, pp. 82-91, Nov. 1996.

[5] E. Gelenbe, X. Mang, and R. Önvural, "Bandwidth Allocation and Call Admission Control in High-Speed Networks," *IEEE Communications Magazine*, vol. 35, no. 5, pp. 122-129, May 1997.

[6] M. Schwartz, Broadband Integrated Networks. Upper Saddle River, New Jersey: Prentice Hall, 1996.