


1999

Reduction of bandwidth requirement by traffic dispersion in ATM networks

Byungjun Ahn
Iowa State University

Follow this and additional works at: <https://lib.dr.iastate.edu/rtd>

 Part of the [Computer Sciences Commons](#), and the [Electrical and Electronics Commons](#)

Recommended Citation

Ahn, Byungjun, "Reduction of bandwidth requirement by traffic dispersion in ATM networks " (1999). *Retrospective Theses and Dissertations*. 12546.
<https://lib.dr.iastate.edu/rtd/12546>

This Dissertation is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Retrospective Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

**A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600**

Reduction of bandwidth requirement by traffic dispersion in ATM networks

by

Byungjun Ahn

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY

Major: Computer Engineering

Major Professor: Douglas W. Jacobson

Iowa State University

Ames, Iowa

1999

Copyright © Byungjun Ahn, 1999. All rights reserved.

UMI Number: 9924697

UMI Microform 9924697
Copyright 1999, by UMI Company. All rights reserved.

**This microform edition is protected against unauthorized
copying under Title 17, United States Code.**

UMI
300 North Zeeb Road
Ann Arbor, MI 48103

**Graduate College
Iowa State University**

**This is to certify that the Doctoral dissertation of
Byungjun Ahn
has met the dissertation requirements of Iowa State University**

Signature was redacted for privacy.

~~Committee Member~~

Signature was redacted for privacy.

~~Committee Member~~

Signature was redacted for privacy.

~~Committee Member~~

Signature was redacted for privacy.

Committee Member

Signature was redacted for privacy.

~~Major Professor~~

Signature was redacted for privacy.

For the Major Program

Signature was redacted for privacy.

~~For the Graduate College~~

TABLE OF CONTENTS

| | |
|--|-----|
| ABSTRACT | vii |
| 1 INTRODUCTION | 1 |
| 2 PRELIMINARIES | 4 |
| VP based ATM networks | 4 |
| VP capacity reservation | 7 |
| VC routing | 9 |
| Equivalent capacity and traffic model | 10 |
| Efficiency of traffic dispersion | 16 |
| 3 PROBLEM STATEMENT | 25 |
| 4 COST EFFECTIVE DISPERSION ROUTING ALGORITHM | 27 |
| 5 SIMULATION RESULTS | 30 |
| 6 ANALYTICAL MODELS | 46 |
| Single path load distribution | 48 |
| Load distribution when CED is used | 59 |
| 7 CONCLUSION | 68 |
| BIBLIOGRAPHY | 70 |

LIST OF TABLES

| | | |
|-----------|--|----|
| Table 5.1 | Mean (σ) of interarrival time at VP_0 | 43 |
| Table 5.2 | Mean (σ) of R_{peak} at VP_0 | 43 |
| Table 5.3 | Mean (σ) of ρ at VP_0 | 44 |
| Table 5.4 | Mean (σ) of b at VP_0 | 45 |
| Table 5.5 | Mean (σ) of b at VP_0 with 4 VP's | 45 |

LIST OF FIGURES

| | | |
|-------------|---|----|
| Figure 2.1 | An example of VP based ATM network | 6 |
| Figure 2.2 | Basic on-off traffic source model | 12 |
| Figure 2.3 | Equivalent capacity for a single source | 15 |
| Figure 2.4 | Equivalent capacity for 5 identical sources | 15 |
| Figure 2.5 | Five identical sources with not enough buffer | 16 |
| Figure 2.6 | Equivalent capacity for twenty homogeneous sources | 17 |
| Figure 2.7 | Generic traffic dispersion technique | 18 |
| Figure 2.8 | Heterogeneous set1 | 22 |
| Figure 2.9 | Heterogeneous set2 | 22 |
| Figure 2.10 | Heterogeneous set1 with two different ρ | 23 |
| Figure 2.11 | Heterogeneous set1 with two different $[\rho, b]$ | 23 |
| Figure 2.12 | Heterogeneous set2 with two different ρ | 24 |
| Figure 2.13 | Heterogeneous set2 with two different $[\rho, b]$ | 24 |
| Figure 4.1 | Cost effective dispersion algorithm | 29 |
| Figure 5.1 | Bandwidth requirement when $[R_{peak} = 8Mbps, \bar{\rho} = 0.1, \bar{b} = 0.5sec]$ | 31 |
| Figure 5.2 | Bandwidth requirement when $[R_{peak} = 8Mbps, \bar{\rho} = 0.1, \bar{b} = 1.0sec]$ | 32 |
| Figure 5.3 | Bandwidth requirement when $[R_{peak} = 8Mbps, \bar{\rho} = 0.2, \bar{b} = 0.5sec]$ | 33 |
| Figure 5.4 | Bandwidth requirement when $[R_{peak} = 8Mbps, \bar{\rho} = 0.2, \bar{b} = 1.0sec]$ | 34 |
| Figure 5.5 | Bandwidth requirement when $[\bar{R}_p = 16Mbps, \bar{\rho} = 0.1, \bar{b} = 0.5sec]$ | 35 |
| Figure 5.6 | Bandwidth requirement when $[\bar{R}_p = 16Mbps, \bar{\rho} = 0.1, \bar{b} = 1.0sec]$ | 36 |
| Figure 5.7 | Bandwidth requirement when $[\bar{R}_p = 16Mbps, \bar{\rho} = 0.2, \bar{b} = 0.5sec]$ | 37 |
| Figure 5.8 | Bandwidth requirement when $[\bar{R}_p = 16Mbps, \bar{\rho} = 0.2, \bar{b} = 1.0sec]$ | 38 |
| Figure 5.9 | Effect of CED <i>coefficient</i> on D_f [8 VP's] | 40 |
| Figure 5.10 | Effect of CED <i>coefficient</i> on D_f [4 VP's] | 41 |

| | | |
|-------------|---|----|
| Figure 5.11 | CDF of interarrival time at VP_0 | 42 |
| Figure 5.12 | CDF of R_{peak} at VP_0 | 43 |
| Figure 5.13 | CDF of ρ at VP_0 | 44 |
| Figure 5.14 | CDF of b at VP_0 | 44 |
| Figure 6.1 | Markov chain for a multiplexer with N VP's | 48 |
| Figure 6.2 | Distribution of \mathcal{M}_N | 52 |
| Figure 6.3 | Distribution of $K\mathcal{S}_N$ | 55 |
| Figure 6.4 | Load distribution estimated by $\mathcal{C}_{(S),N}$ analytical model | 58 |
| Figure 6.5 | Accuracy of $\mathcal{C}_{(S),N}$ analytical model | 58 |
| Figure 6.6 | Time-varying VP loads | 60 |
| Figure 6.7 | Distribution of ρ seen by individual VP | 61 |
| Figure 6.8 | Distribution of interarrival time at a VP | 62 |
| Figure 6.9 | Gaussian approximation of R_{sap} | 63 |
| Figure 6.10 | R_{sap} distribution with tailored mean and σ^2 | 64 |
| Figure 6.11 | Accuracy of \tilde{D}_f estimator | 64 |
| Figure 6.12 | Distributions of \mathcal{M}_N at Single Abstract Path | 66 |
| Figure 6.13 | Distributions of $K\mathcal{S}_N$ at Single Abstract Path | 66 |
| Figure 6.14 | Load distributions at Single Abstract Path | 67 |

ABSTRACT

The problem of bandwidth allocation and routing in Virtual Path (VP) based Asynchronous Transfer Mode (ATM) networks was 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 is 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 a bursty and heterogeneous multimedia traffic.

First, the effect of traffic dispersion was explored to achieve more statistical gain. Through this study, it was discovered how the effect of traffic dispersion varies with different traffic characteristics and the number of paths. An 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 the CBP can be significantly reduced by CED.

Next, this study provides 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. Then analytical models have been formulated. The cost effective design and engineering of VP networks requires accurate and tractable mathematical models which capture the important statistical properties of traffic. This study also revealed that the load distribution estimated by equivalent capacity follows Gaussian distribution which is the sum of two jointly Gaussian random variables. For the analysis of load distribution when CED is used, we simplified multiple paths as identical paths using the idea of Approximation by Single Abstract Path (ASAP), and approximated the characteristics of the traffic seen by individual

VP. The developed analytical models and approximations were validated in the sense that they agreed with simulation results.

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.). As new bursty and heterogeneous traffic classes are added, network traffic changes more dynamically and needs a much more complex model to capture the important statistical properties of it. Thus efficient management of network resources (e.g., bandwidth allocation, etc.) becomes a rigorously difficult task for network engineers. We have to confront with many new types of traffic classes whose traffic models are currently either unknown or poorly understood.

Asynchronous Transfer Mode (ATM) is the transfer mode for implementing Broadband Integrated Services Digital Networks (B-ISDNs) and other high-speed networks. The term “*transfer*” comprises both transmission and switching aspects, so a transfer mode is a specific way of transmitting and switching information in a network. ATM provides a unified interface which 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. When a deterministic bandwidth allocation scheme is used, bandwidth for each connection is allocated in a straightforward way. As an example, for each connection, peak bit rate of the connection can be allocated. The advantage of this scheme lies in its simplicity of call admission control (CAC). This is because only knowledge of the peak bit rate of the traffic source is required for CAC. The new connection is accepted if the sum of the peak rates of all the existing connections and the new connection does not exceed link capacity. As long as traffic sources transmit cells at their peak rates,

this scheme seems to be very efficient. However, when a deterministic bandwidth allocation scheme is used for bursty traffic sources, large amount of bandwidth can be wasted, particularly for those with large peak to average bit rate ratios. For bursty traffic sources, statistical multiplexing is desirable to achieve high utilization of network resources. Since bursty traffic sources are not likely to transmit cells at their peak rates simultaneously all together, when statistical multiplexing is used, bandwidth for each connection is allocated less than its peak rate, but necessarily greater than its average bit rate. But accurate estimation of both statistical bandwidth requirement of an individual connection and the aggregate bandwidth usage of connections which are multiplexed to a given link, pose a formidable challenge. This is because the statistical bandwidth of a connection depends not only on its own stochastic characteristics, but also strongly on the characteristics of existing connections in the link. Another difficulty in designing an efficient algorithm for the statistical bandwidth allocation is that decisions must be made on the fly.

Most statistical bandwidth allocation schemes are based on the well-known concept of effective bandwidth which has been studied extensively. Section 2 outlines fundamentals of equivalent capacity (or effective bandwidth used in this study), based on the work by Guérin *et al.* [1]. The concept of the equivalent capacity is widely used in CAC schemes to design admission criteria because of its efficient approximations for the problem of bandwidth allocation and simplicity amenable to real-time computations. 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]. By using traffic dispersion, a burst is divided into many number of sub-bursts, which are transmitted in parallel through multiple paths and are resequenced at the destination. However, the impact of traffic dispersion on the effective capacity has not been studied well in the literature. There has been no thorough report on efficient traffic dispersion algorithms which are generally applicable in most of network traffic situations. This is due to the lack of accurate and computationally amenable traffic model.

As an efficient way to facilitate the network management, Virtual Path (VP) concept has been proposed. A VP is a direct logical connection between two end nodes. More than one VP's can be established between a pair of nodes. A logical VP network can be defined by viewing VP end points

as nodes and VP's as links. A lot of different logical VP networks can be configured dynamically for a given physical ATM network topology. Ultimately, network providers will design economic VP networks to meet both cell and call level QoS requirements (where the call level QoS requirement is that the blocking probability be below some level).

Traffic control and resource management are simplified in VP based networks. However, *a priori* reservation of resources on VP's also reduces the statistical multiplexing gain of the network, resulting in increased Call Blocking Probability (CBP). The focus of this study is on how to reduce CBP (or how to improve bandwidth efficiency for a given CBP requirement) by effective bandwidth allocation and routing algorithms, when link capacity is reserved on VP's. Equivalent capacity concept is used to calculate the required bandwidth by the call. Each call is represented as a bursty and heterogeneous multimedia traffic. First, we explore the effect of the traffic dispersion to achieve more statistical gain. Throughout this study, it is discovered how the effect of traffic dispersion changes with different traffic characteristics and the number of paths. Efficient routing algorithms including a cost effective traffic dispersion algorithm are designed. Simulation study shows that call blocking probability (CBP) can be significantly reduced by the algorithm. Next, this study provides 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. Then analytical models are formulated, which captures the actual load distribution precisely. The cost effective design and engineering of VP networks requires accurate and tractable mathematical models which capture the important statistical properties of traffic. This study reveals that the load distribution estimated by equivalent capacity follows a Gaussian distribution which is the sum of two jointly Gaussian random variables. For the analysis of load distribution when cost effective traffic dispersion is used, we simplify multiple paths as identical paths using the idea of Approximation by Single Abstract Path (ASAP), and approximate the characteristics of the traffic seen by individual VP's as a result of the dispersion. The developed analytical models and approximations are proven to be correct when compared with simulation results.

The organization of this paper is as follows. Chapter 2 provides a brief introduction to related work. In chapter 3, the objectives and assumptions for this study are illustrated. A cost-effective traffic dispersion algorithm is presented in Chapter 4, while simulation design and results are discussed in Chapter 5. Chapter 6 is devoted to analytical model. Finally, Chapter 7 summarizes the findings of this work and outlines possible future work.

2 PRELIMINARIES

VP based ATM networks

An ATM network is essentially constructed with nodes, physical links such as transmission lines, and terminals. Before a user starts transmitting cells over an ATM network, a connection has to be established. This is done in a call setup procedure. The main objective of this procedure is to establish a path between the sender and the receiver; this path, a "*Virtual Circuit*" (VC) in ATM terms, may be routed through one or more ATM switches. On each of these ATM switches, resources (e.g., bandwidth) have to be allocated to the new VC. Cells are transferred along the route assigned to the VC to which they belong. In this network, several node functions are needed. First, each node needs to recognize the outgoing link or terminal to which incoming cells should be sent. They also must determine the route for the VC upon call request and rewrite routing tables to setup a route. If sufficient bandwidth can not be reserved in the network for the requested call, the call connection should be refused to prevent network congestion. By call blocking, specified transmission qualities such as cell delay and cell loss rate are assured for all connected VC's. Thus, the decision to refuse or accept the call at VC setup procedure is another necessary node function although the definition of bandwidth for bursty traffic (i.e., statistical bandwidth) is a remaining problem that must be solved. As an effective means to construct an ATM network, the concept called "*Virtual Path*" (VP) has been proposed. The VP is a concept to simplify and eliminate some of these node functions. In particular, it reduces processing loads generated by a VC setup procedure. The benefits provided by the concept also include simplified node structure, reduced control of routing and bandwidth. The VP is an information transport path defined as follows.

- A VP is a logical direct link between two nodes and accommodates a number of VC's simultaneously.
- A predefined route is allocated for each VP in the physical facilities network.

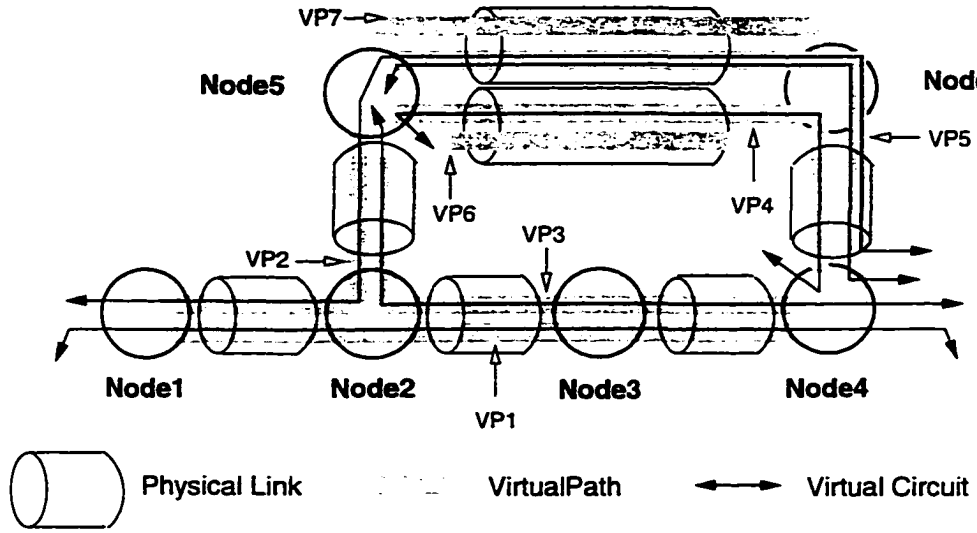
- Each VP has a bandwidth, in other words, “*capacity*” which defines the upper limit for the total VC bandwidth carried by it.
- VP’s are multiplexed on physical transmission links in a cell multiplexing manner. In other words, transmission facilities are shared by a number of VP’s.

Fig. 2.1 shows an example of VP based ATM network. Note that more than one VP’s can be established between a pair of nodes. A number of different logical VP networks can be configured dynamically for a given physical ATM network topology.

By using VP’s with the features mentioned earlier, the required node functions are effectively simplified. First, at VC setup, it becomes unnecessary to rewrite the routing table of the transit nodes such as node2 and node3 for VP1 in Fig. 2.1(a). This is made possible by Virtual Path Identifier (VPI) in each cell. At the transit nodes through which the VP passes, an outgoing link for arrived cells can be recognized by comparing their VPI’s with the routing table. Since the routing table is concerned with the VP and not associated with each call, rewriting of the table is not necessary at VC setup. Routing procedure at VC setup is also eliminated at the transit nodes because this is done by selecting the most appropriate VP from the end nodes terminating the VP. Additionally, the transit nodes are free from the bandwidth allocation process at VC setup. This can be carried out by comparing the bandwidth of the requested call to the unused bandwidth of the VP at the end nodes. Thus, by reserving resources on VP’s, all VC setup functions can be eliminated from the transit nodes in the VP based ATM network.

Also, VP’s can be used to segregate traffic types which require different QoS’s, and to facilitate aggregation with similar performance requirements. One rationale for using VP’s for traffic control is to separate different traffic types in order to prevent them from interacting or interfering with each other. For example, different types of traffic (e.g., voice and data) may have very different traffic characteristics and QoS requirements: the question of how to multiplex two or more diverse traffic classes while providing different QoS requirements at a switch, is a very complicated, open problem. The problem can be simplified if different types of traffic are separated by assigning a VP with dedicated resources to each type of traffic. Usually, VP based ATM networks are designed such that each VP carries one type of traffic with a specific QoS requirement.

The efficient design and engineering of VP based ATM networks require three design issues to be addressed: VP capacity reservation, VC routing, and the topology of VP network (i.e., the layout of VP’s on a given ATM network). In [3], Gerstel *et al.* address the problem of designing a layout of VP’s on a given ATM network. In order to simplify the problem, authors of [3] make several assumptions.



(a) Schematic illustration of VP's

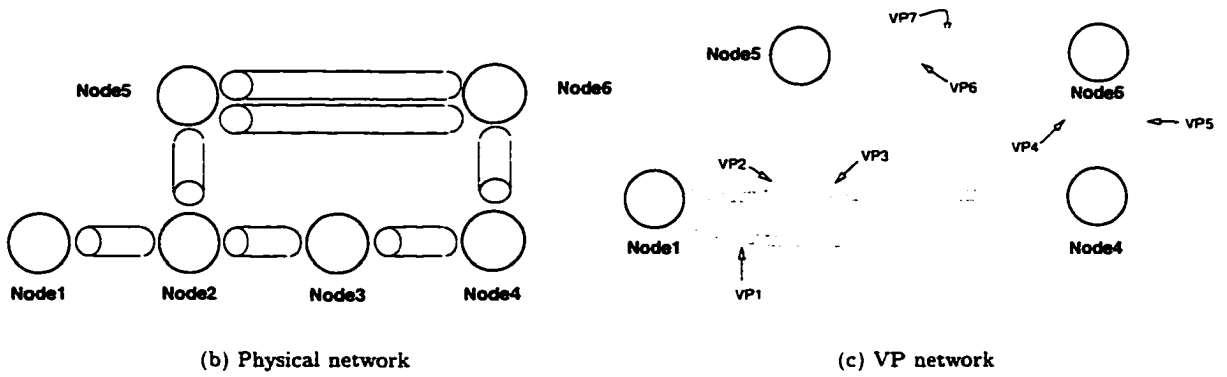


Figure 2.1 An example of VP based ATM network

The load on a link is defined as the number of VP's that include the link in their paths. And only a single route between pairs of nodes is allowed. In our study, however, multiple VP's between any pair of nodes are considered. And we define the link load as the aggregate bandwidth of all VP's passing through the link. The bandwidth dedicated to each VP (i.e., VP capacity) differs from VP to VP.

In this study, we do not explicitly address the topology design problem. Instead, our approach to this problem is to 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. Furthermore this assumption makes it possible to restrict all candidate VC routes to consist of single VP (i.e., direct routing policy). In direct routing, calls are only allowed to be routed via a direct VP between a source node and a destination node.

One of primary objectives of this study is to improve network utilization through optimal VP bandwidth reservation when multipath VC routing algorithms are used. We also show that the optimal number of VP's between a pair of nodes can be estimated by a deterministic way. In the following two sections, we briefly review related previous work on the VP capacity allocation and VC routing.

VP capacity reservation

In the literature, two VP capacity reservation strategies have been used: a deterministic strategy and a statistical strategy [4].

The deterministic strategy, advocated by [5], reserves separate link capacity for each VP passing through the link. This approach treats a VP as if it were a physical link. A certain amount of buffer space at the source node and bandwidth at each physical link on this VP are dedicated to this VP. Since there is no statistical multiplexing performed among VP's at the link level, the sum of the reserved VP bandwidths on a link is not permitted to exceed the total capacity of this link. However, statistical cell multiplexing is still performed among VC's within a VP. Therefore, it is possible for the offered peak rate to a VP to exceed its allocated bandwidth over a short period of time. In such a case, cells that cannot be transmitted are buffered at the source node of the VP. Since the instantaneous transmission rate of a VP is limited to its reserved bandwidth, the buffer space required and the queueing delays incurred at transit nodes/links of the VP are very small, and are assumed to be negligible [5]. By making use of the fluid approximation technique [6] and identical traffic assumption, Gupta *et al.*, in [5], could define the capacity of a VP as the maximum number of VC's that the VP can carry.

The study by Ohta *et al.* [7] is rather different from [5]. In their study, VP capacities are dynamically increased/decreased on demand in units of predetermined step size, instead of reserving capacity on

VP's for fixed length periods. By assuming that every call is homogeneous and that the traffic offered to each VP is identical, the authors study the effects of the step size on the trade-off between control cost and transmission cost. No statistical multiplexing effect is considered in allocating the bandwidth for a call or a VP, (i.e., the bandwidth is allocated deterministically).

Obviously, both the identical traffic assumption and the fluid approximation technique which is used to estimate the bandwidth requirement for a VC, can not be used in our study. The fluid approximation technique is known to be very conservative when the number of VC's multiplexed onto a VP is large. In order to overcome this conservativeness, a stationary approximation technique has been proposed by Guérin *et al.* [1]. Instead of estimating the the bandwidth requirement for an individual VC, the stationary approximation technique estimates the aggregate bandwidth requirement of all VC's on a VP. Throughout this study, the approach proposed by Guérin *et al.* [1] is used to estimate bandwidth requirement.

The statistical strategy, inspired by [4], allows statistical multiplexing of cells from different VP's on a physical link. In this approach, instead of explicitly reserving buffer space and link bandwidth for VP's, each VP is allocated a dedicated number of VC's, each with a guaranteed QoS, such that call setup processing can still be done locally at the source node so long as the number of existing VC's within the VP is less than this number. The actual reservation is done by pretending as if each VP has accepted as many as the dedicated number of VC's into the network. The advantage of the statistical approach is that it provides better statistical cell multiplexing gain than the deterministic strategy. However, this advantage is offset by the fact that when some connections are routed over multiple VP's, a more stringent QoS guarantee needs to be provided by each component VP. Two problems can be identified when implementing the statistical strategy. First, the traffic characteristics of a VC may change as the traffic travels along the path: these changes must be characterized. Second, we must determine how to assign the end-to-end QoS requirement to each link such that the overall end-to-end QoS can not be satisfied.

Beside these two problems, a major drawback of this statistical VP capacity reservation strategy arises from the assumptions which had to be made. By assuming that the traffic characteristic of every call offered to a VP is identical, a dedicated number of VC's for each VP could be allocated. Apparently, this approach would not work when heterogeneous traffic sources are offered to a VP, which have the same QoS requirement (e.g., cell loss ratio) but different traffic characteristics (e.g., peak rate and burstiness).

VC routing

Traffic control and resource management are simplified in VP based networks. However, *a priori* reservation of resources on VP's also reduces the statistical multiplexing gains of the network, resulting in an increased Call Blocking Probability (CBP). Earlier research in VP networks, [8] for example, has focused on the reduction of CBP (or increasing bandwidth efficiency) in a homogeneous network. Thereafter, the study of routing algorithms to reduce CBP (i.e., to increase network efficiency) has been an active research issue in VP based ATM networks. With the assumption of identical bandwidth requirements for each call and no statistical multiplexing effect among calls, the routing of VC's is similar to that of connections in conventional circuit switched networks. Thus, many dynamic routing algorithms studied in the conventional networks have been applied to VP networks. Two well-known types of routing algorithms are Least Loaded Path (LLP) based and Markov Decision Process (MDP) based algorithms. The LLP method routes the alternate calls to the path with the maximum number of free circuits while MDP method formulates the routing problem as Markov Decision Process. Gupta *et al.* [5] studied the LLP approach, and Hwang *et al.* [9] studied MDP. Both of these studies used simulation. Wong *et al.* [10] also studied a LLP based routing algorithm called Maximum Free Circuit routing.

Our study differs from this previous work in the following respects. The assumption of identical bandwidth requirement for each call was crucial in all previous work. However, since we consider heterogeneous traffic multiplexed onto a VP, this assumption cannot be justified. Instead of estimating the the bandwidth requirement for individual VC, the approach proposed by Guérin *et al.* [1] is used in our study to estimate the aggregate bandwidth requirement of all VC's on a VP.

Second, we consider direct routing (or single hop routing) algorithms. In direct routing, calls are only allowed to be routed via a direct VP between a source node and a destination node. In [9, 10], a direct path and a few two-hop alternate paths are considered for each node pair. In our study, however, it is assumed that any pair of nodes in the network has one or more VP's for each different class of QoS requirement. We provide a deterministic way of estimating the optimal number of VP's between a pair of nodes. The focus of this study is on how to reduce CBP (or how to improve bandwidth efficiency for a given CBP requirement) when establishing multiple single-hop VP's is possible between a pair of nodes. Furthermore, we explore the effect of the traffic dispersion, which 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]. We first design efficient routing algorithms including a cost effective traffic dispersion algorithm, and show that the network blocking probability

can be significantly reduced by the routing algorithm. We then examine VP capacity reservation and VP network design strategies.

Finally, we use a deterministic VP capacity reservation strategy. With the assumption of identical bandwidth requirement for each call in a VP, Hwang *et al.* [9] considered a statistical VP capacity reservation strategy, in addition to studying a deterministic strategy. When capacity is reserved on VP's, it may be desirable to dynamically adjust the allocation to improve link bandwidth utilization as well as to adapt to dynamic changes in network traffic flows. 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.

Equivalent capacity and traffic model

Bandwidth allocation is handled by CAC, and estimating the accurate statistical bandwidth required is of most importance in any CAC strategy. Estimating the value of the statistical bandwidth for an incoming call request must address the following issues:

- the QoS requirements of the new connection must be guaranteed.
- the QoS provided to connections established previously must not be degraded to unacceptable levels when they are multiplexed with the new connection.

Accordingly, the statistical bandwidth of a connection depends not only on its own stochastic characteristics, but also strongly on the characteristics of existing connections in the network. There has been much research on methods to evaluate the required bandwidth of aggregate sources of ATM traffic, and many solutions have been suggested. [11] and [12] provide a handy tutorial on the bandwidth allocation and call admission control in high-speed networks.

Perros and Elsayed, in [11], classified CAC schemes into the following groups based on the underlying principle that was used to develop the scheme:

- equivalent capacity
- heavy traffic approximation
- upper bounds of the cell loss probability

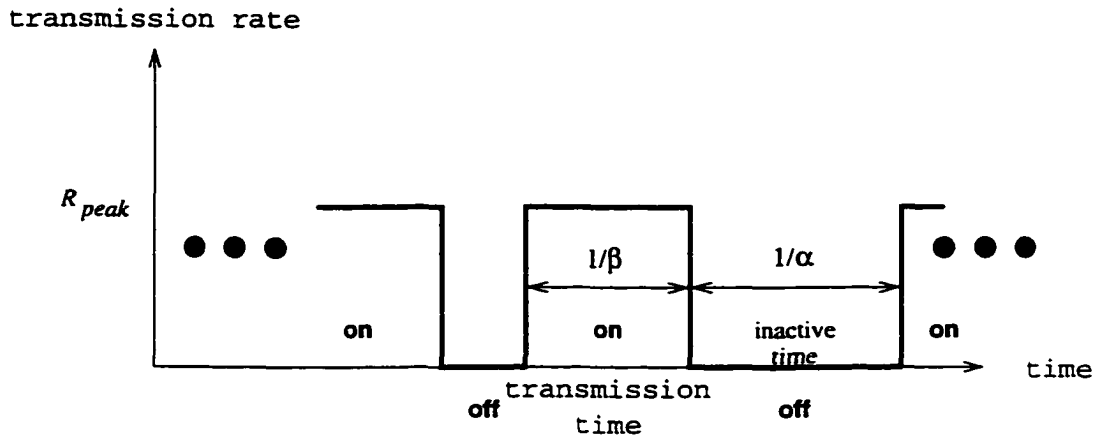
- fast buffer/bandwidth allocation
- time windows.

A numerical performance comparison of some call admission schemes is also presented in [11]. It shows that the equivalent capacity scheme, proposed by Guérin *et al.* [1], outperforms other schemes in many cases.

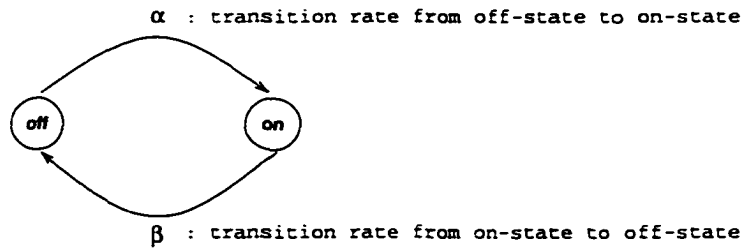
Equivalent capacity provides a unified metric to represent the effective bandwidth of a connection as well as the effective aggregated load on network links at any given time. In other words, equivalent capacity can be defined as the bandwidth requirement of a single or multiplexed connections on the basis of their statistical characteristics. Equivalent capacities (or effective bandwidths) for a variety of traffic source models have been obtained by a host of research groups [1, 13, 14, 15, 16, 17]. In [13, 14] equivalent capacities for multiple Markov modulated fluid-flow sources sharing a statistical multiplexer with a large buffer is obtained. Gibbens and Hunt have extended the fluid-flow analysis of Anick *et al.* [6] to show that the notation of additive equivalent capacity exists for a set of heterogeneous on-off sources jointly accessing a channel [14]. Kesidis *et al.* [15] have also shown the equivalent capacity expression for the sources represented by the Markov-modulated Poisson process (MMPP) using techniques from large-deviation theory. The theory is being used widely to study the probabilistic behavior of large excursions or deviations of random variables from their mean values. The theory is particularly useful in studying tail probabilities of random variables, and in obtaining estimates of their asymptotic behavior. Many authors explored these large-deviation techniques to estimate the effective bandwidth requirement (i.e., equivalent capacity) [18, 19, 20, 21, 22, 23].

The concept of equivalent capacity can be generalized to a variety of traffic source models, including those such as the Markov-modulated models which 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. 2.2. 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 [24]. 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.

An on-off traffic source requires at least three parameters (i.e., traffic descriptors) to represent it:



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



(b) Markov fluid-flow model

Figure 2.2 Basic on-off traffic source model

- R_{peak} : connection's peak rate
- $\frac{1}{\beta}$: mean burst length, and
- $\frac{1}{\alpha}$: mean idle time.

Or alternatively.

- R_{peak} : connection's peak rate
- $\rho = \frac{\alpha}{\alpha+\beta}$: utilization of the connection (i.e., fraction of time the source is active)
- b : mean of the burst period.

Heterogeneous sources would have different values for the three parameters. For homogeneous sources they would be the same. In this study we assume that the statistics of both intervals are exponential and independent from each other. In [16] an effective bandwidth was derived for heterogeneous on-off sources where on periods and off periods are dependent and generally distributed, in contrast to the widely used assumptions of independent and exponentially or geometrically distributed periods. Kulkarni *et al.* [25] considered the equivalent capacity vector for two-priority on-off source.

Guérin *et al.* [1] have proposed the following expression as a good approximate measure of equivalent capacity \hat{C} :

$$\hat{C} = \min[\hat{C}_{(F)}, \hat{C}_{(S)}] \quad (2.1)$$

where $\hat{C}_{(F)}$ is a fluid-flow approximation given by Eq.(2.2) and $\hat{C}_{(S)}$ is a stationary approximation given by Eq.(2.3).

The fluid-flow approximation, $\hat{C}_{(F)}$, 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, as shown by Eq.(2.2).

$$\hat{C}_{(F)} = \sum_{i=1}^N \frac{\gamma_i - x + \sqrt{(\gamma_i - x)^2 + 4x\rho_i\gamma_i}}{2b_i(1 - \rho_i) \ln \frac{1}{\epsilon}} \quad (2.2)$$

where $\gamma_i = b_i(1 - \rho_i)R_{peak}^i \ln \frac{1}{\epsilon}$, and x is the given buffer size in bits.

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 ϵ (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, given by Eq.(2.3), provides reasonably accurate estimate, which approximates the distribution of the stationary bit rate on a link.

$$\hat{C}_{(S)} \approx m + K\sigma, \quad \text{with} \quad K = \sqrt{-2 \ln \epsilon - \ln(2\pi)} \quad (2.3)$$

where, m is the mean aggregate bit rate ($m = \sum_{i=1}^N m_i$), and
 σ is the standard deviation of the aggregate bit rate ($\sigma^2 = \sum_{i=1}^N \sigma_i^2$).

As both approximations overestimate the actual value of the equivalent capacity for different range of connections characteristics, the equivalent capacity \hat{C} is taken to be the minimum of $\hat{C}_{(F)}$ and $\hat{C}_{(S)}$ to predict the relatively accurate equivalent capacity of connections.

In our study, Fig. 2.3 - Fig. 2.6 are obtained from simulations and are in agreement with [1]. Fig. 2.3 shows that, as expected, the fluid-flow approximation estimates the accurate equivalent capacity for a single source.

Fig. 2.4 illustrates the behavior of the equivalent capacity as a function of source utilization ρ when a small number, 5, of homogeneous traffic sources are multiplexed, each with 4Mbps peak rate and 0.1 sec. mean burst period. In this example, flow approximation captures the required capacity reasonably well. On the other hand, stationary approximation results in a substantial overestimate of the required capacity. This is because, as mentioned above, the basic assumption of stationary approximation was the large number of traffic sources multiplexed together. Five traffic sources are not enough for stationary approximation to work properly. From simulation results, it was learned that at least more than 10 are required for the fair estimate. For more than 20 sources, stationary approximation works well.

Now we evaluate the equivalent capacity when buffer size is not enough. Note that fluid-flow approximation counts on the size of buffer while stationary approximation does not. With the same buffer size of 3 Mbit as used in Fig. 2.4, if we increase either peak rate or mean burst period, we get the equivalent capacity as illustrated in Fig. 2.5. As either peak rate or mean burst period is increased by 10 times, more buffer capacity would be needed to guarantee the claimed cell level QoS (i.e., Cell

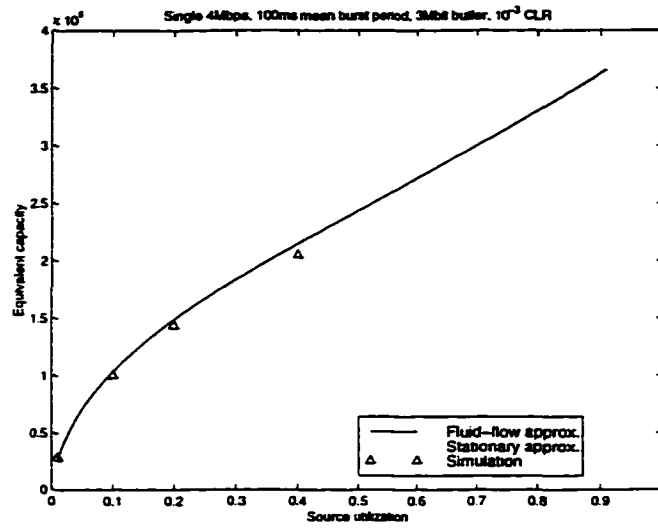


Figure 2.3 Equivalent capacity for a single source

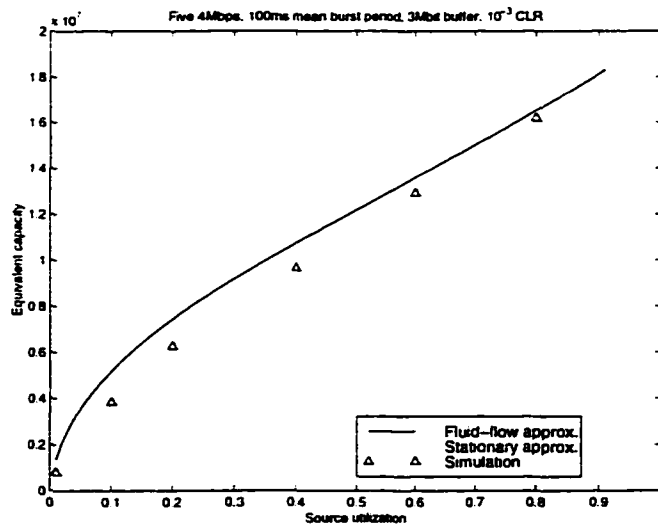


Figure 2.4 Equivalent capacity for 5 identical sources

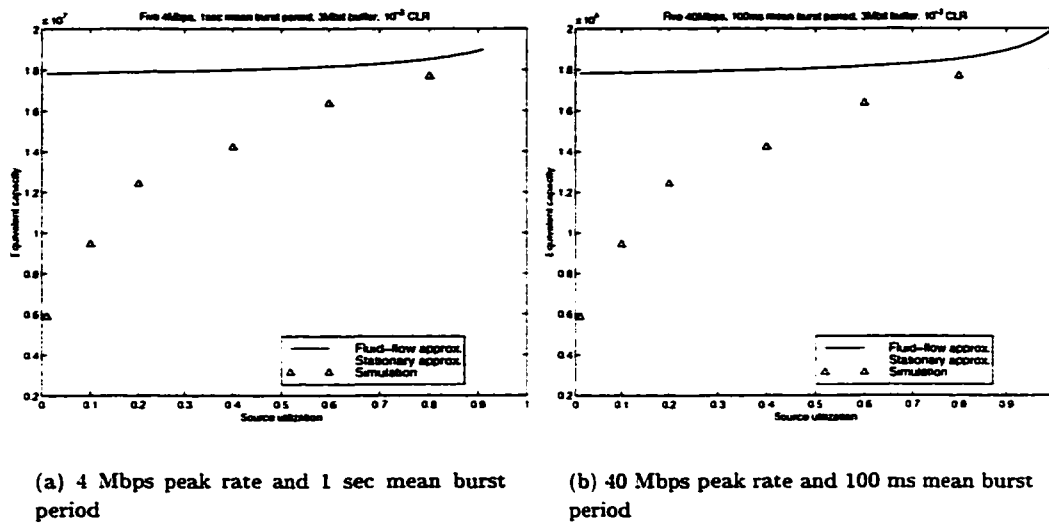


Figure 2.5 Five identical sources with not enough buffer

Loss Ratio). In practice, however, we cannot change the buffer size of the multiplexer dynamically. Furthermore, the larger the buffer size, the longer the node delay. Instead, the equivalent capacity method counters increased peak rate and mean burst time by estimating the larger required capacity. In Fig. 2.5, stationary approximation provides better estimate at low source utilization, while fluid-flow approximation is again more accurate at high source utilization.

Fig. 2.5(b) demonstrates that, if two traffic have the same $R_{peak} \times b$ (i.e., peak rate mean burst period product), behavior of the equivalent capacity of those two traffic are analogous. In this example, Fig. 2.5(b) is analogous to Fig. 2.5(a). Fig. 2.6(a) considers the multiplexing of larger numbers, 20, of homogeneous traffic sources, each with 10 Mbps peak rate and 100 ms mean burst period. It exhibits that, as the number of connections increased, the stationary approximation performs better, although it also overestimates. Fig. 2.6(b) illustrates the behavior of equivalent capacity for the traffic with longer mean burst period. Of importance here is the fact that the fluid-flow and stationary approximations complement each other over different ranges of traffic characteristics.

Efficiency of traffic dispersion

One of the major attractive feature of ATM network is the use of statistical multiplexing, which allows several sources to share the capacity of a link statistically. This means that the demand for

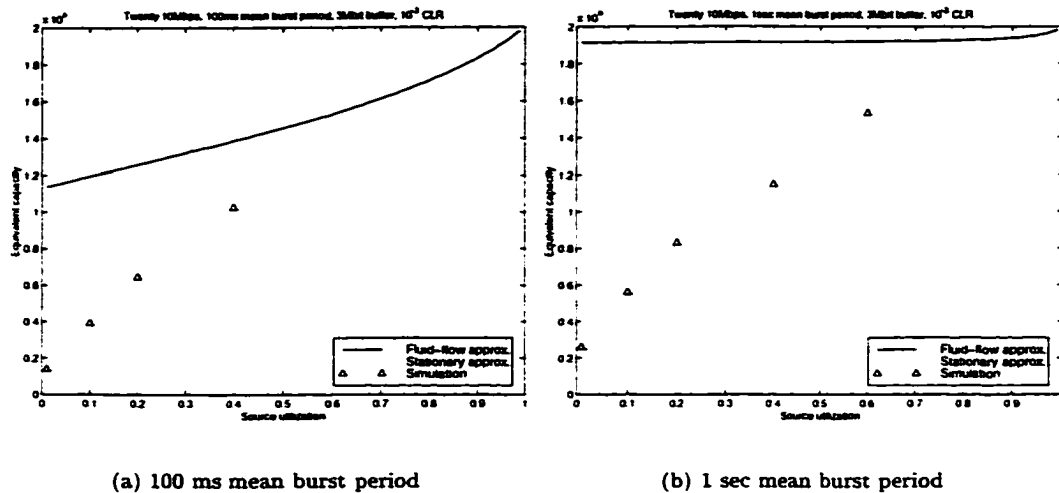
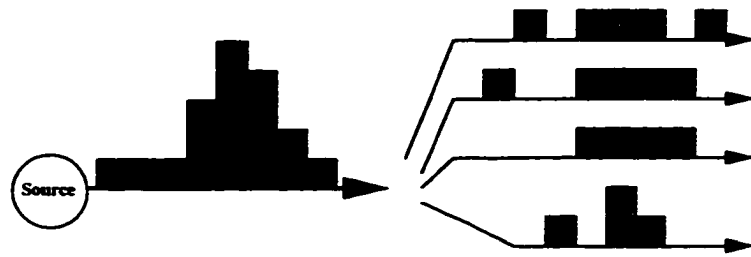


Figure 2.6 Equivalent capacity for twenty homogeneous sources

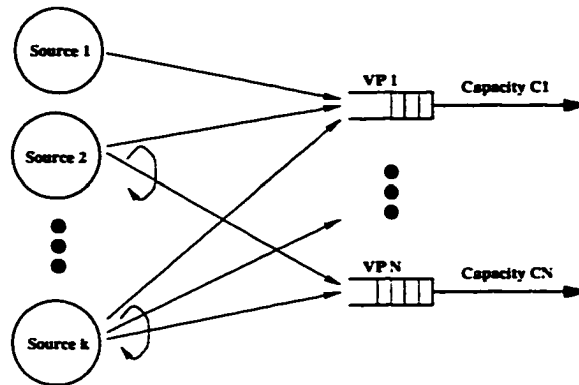
capacity may sometimes exceed what is available, and cells may be lost. Such a problem of statistical multiplexing is aggravated when the traffic arrives in bursts. Recent studies [26, 27] indicate that the data traffic exhibits pronounced correlation with a high variance over long time periods. Keeping the loss probability at a tolerable level for such traffic would require unreasonably low utilization of the network capacity. The problem might be handled at a source node by spreading the cells in time (shaping). However, when the traffic is generated at a high rate in long bursts, the delay introduced by the shaping may become unacceptable.

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. 2.7 shows an illustration of spatial traffic dispersion and a model of a one stage multiplexer where k sources generate traffic which 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) Illustration of spatial traffic dispersion



(b) One-stage multiplexer

Figure 2.7 Generic traffic dispersion technique

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 in case the optimal path for some reason becomes congested. 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.

Traffic dispersion is a topic gaining interest, and much work has recently been done in the field.

Most of the work has concentrated on the evaluation of the gain in statistical sharing for the dispersion of various granularity. [28] provides a good survey on traffic dispersion. The work in [29] gives an introduction to the method and investigates how dispersion affects the queuing behavior in a one-stage multiplexer. This study is performed for the fixed number of sources, and all sources are dispersed over the same number of paths (i.e., same dispersion factor). The queuing behavior is demonstrated as a function of the given dispersion factor N . Calculations and simulations show that cyclic dispersion of ATM cells (as shown in Fig. 2.7(a)), generated by a two-state Markov chain, reduces the mean queue size as well as the overall queue size. Simulation results indicate that cyclic dispersion is superior to dispersion of longer sequences of cells, regarding the queuing behavior. Also, it turns out that the mean queue size starts to level out when dispersing the traffic over five paths or more. The work in [2] investigates how dispersion affects the equivalent capacity [1], of a connection in a one-stage multiplexer. It presents for what values of source peak rate, burstiness and peak-to-link ratio, spatial traffic dispersion is useful. However, the results are obtained when a single source or multiple identical traffic sources are multiplexed at an one-stage multiplexer. In their work, network link status (e.g., utilization) is not considered, which changes dynamically as connections are established and torn down. The traffic dispersion strategy used in their work is dividing the source peak rate evenly by the number of links involved. Their results indicate that dispersion decreases the equivalent capacity in general, and in particular for sources with high peak-to-mean ratio and high peak-to-link ratio. These are actually the sources that are most difficult to handle in terms of statistical multiplexing. A call admission control (CAC) scheme with traffic dispersion is presented in [30, 31]. The CAC scheme works as follows. At call request, the source declares its peak rate. A dispersion factor (i.e., the degree of dispersion) is decided for the source, and the source mean rate is estimated. The dispersion factor is chosen between 1 and 10 according to the linear function of source peak rate. The source is then approximated by a Poisson process with bulk arrivals, where the bulk size corresponds to the peak rate divided by the number of paths over which the traffic from the source is spread. The capacity required for the source on one link is calculated with respect to an assigned mean queue size. The capacity is adjusted according to the current measured mean queue size in the multiplexer buffer, and if there is enough capacity available on the link, capacity for the source is allocated and the source starts sending. At call setup, the capacity to be allocated depends on the current measured mean queue size. That is, if measured mean queue size is larger than the assigned value, the allocated capacity needs to be increased, and if the measured mean queue size is smaller than the assigned value, the allocated capacity may be decreased. This makes the capacity allocation and call acceptance decisions depend strongly on the mean queue size

measurements.

Yet another implementation of the traffic dispersion idea is the string mode protocol, presented by Déjean *et al.* [32]. The principle concept of the string mode is to chop bursts into smaller sub-bursts, called strings, and distribute them onto a number of parallel links (virtual paths). At any time, data from a source are transmitted on only one virtual path, so that sharing at the transmitting endpoint can be viewed as occurring serially in time. Routing of strings is determined during connection setup. And individual ATM virtual paths are not allocated at connection setup time, but are determined on the fly when each string is inserted into the network. In their work, the string length was determined by a fixed function of the peak cell rate of source traffic. Each string is distributed onto a number of links either in a random order according to an uniform distribution function or in a cyclic manner (round-robin). In either case, the distribution is done without considering the load status of the VP's. They performed computer simulations to evaluate the string mode protocol, which was intended to investigate the impact of string length, peak bit rate, and allocated number of links with respect to the cell loss probability in the buffer of the source node. The traffic characteristic of each source was assumed to be identical.

Chowdhury analyses the delay introduced by the resequencing function, which is needed when packets are distributed over multiple parallel links [33]. The subject of resequencing is also discussed by Jean-Marie and Gün [34].

Throughout the work done so far, there is no thorough investigation of routing algorithm capable of finding an optimum number of paths nor is there an accurate dispersion strategy for various traffic characteristics. These are issues on which we focus in this study.

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. In order to obtain the same load on a path with as without dispersion, we assume that the path, instead of carrying the traffic from a given number of independent original sources, now carries the traffic from N times as many independent dispersed sources. That is, one path carries fractions of the traffic from each of N times as many original and independent sources. This justifies the independence criterion used in the capacity calculation.

In this study, we refer to traffic dispersion as cyclic spreading of the cells from a source over the available paths. 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_{cell}^{peak} cells/unit-time, cyclic dispersion of cells corresponds to reducing the peak rate of a source on each of the paths to $\frac{R_{cell}^{peak}}{N}$ cells/unit-time, while source utilization ρ and mean burst period b are kept same.

A number of different cases are examined, and presented in Figs 2.8 - 2.13, which show, for each value of N , the aggregated equivalent capacity of a given set of heterogeneous traffic sources. The aggregated equivalent capacities were computed using Eq.(2.2) and Eq.(2.3) for various buffer sizes and a desired cell loss ratio (CLR) of 10^{-3} .

Essentially two different sets of heterogeneous traffic sources are used as the base set, and several variant sets are produced from each base set. The impact of different peak rates is first considered in Fig. 2.8, where 20 sources, all with a source utilization ρ of 0.1 and a mean burst period b of 100 ms but uniformly distributed peak rates ranging from 1 Mbps to 20 Mbps, are multiplexed onto a single path.

The next base set of heterogeneous sources, shown in Fig. 2.9, considers a situation where a number of relatively low-speed sources are multiplexed with a high-speed bursty sources. The multiplexed sources consist of 18 identical sources with a peak rate of 1 Mbps and two sources with a higher peak rate of 20 Mbps. This example examines the impact of very different peak rates in computing the required capacity allocation.

Other variant sets are originated from these two base sets to investigate some aspects of the interactions between sources with different mean burst periods and source utilization. Fig. 2.11 and Fig. 2.13, for example, attempt to study the potential impact of bursty sources on bandwidth requirements of Fig. 2.10 and Fig. 2.12 respectively.

In general, equivalent capacity decreases as N grows. However, 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 further 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.

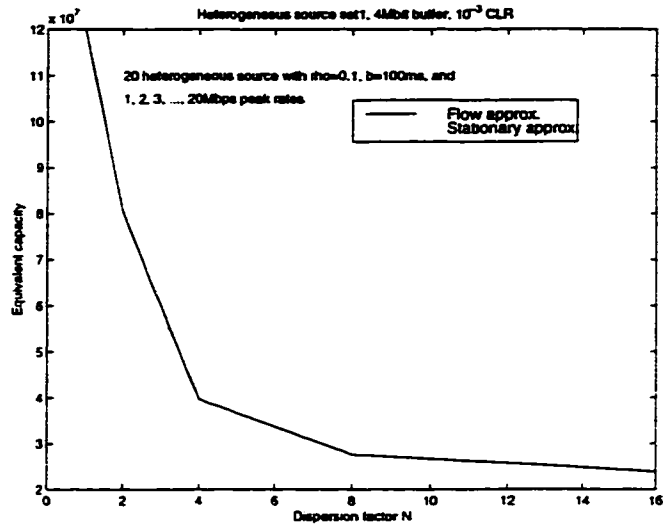


Figure 2.8 Heterogeneous set1

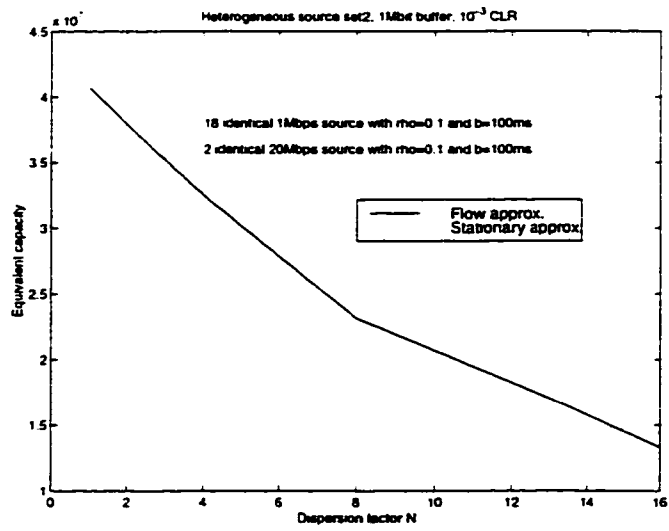


Figure 2.9 Heterogeneous set2

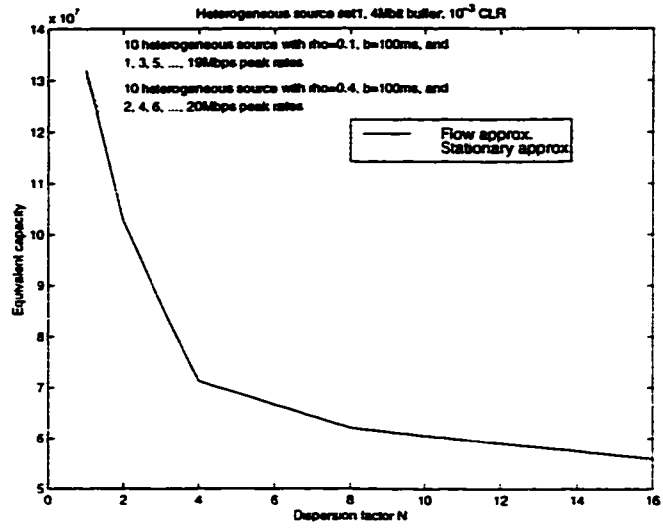


Figure 2.10 Heterogeneous set1 with two different ρ

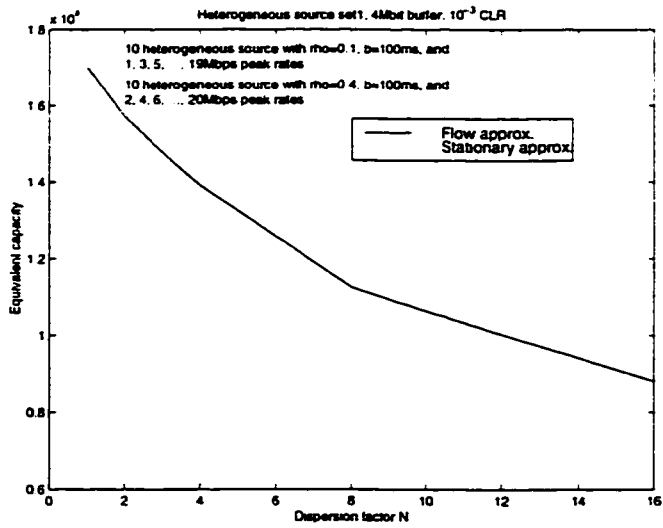


Figure 2.11 Heterogeneous set1 with two different $[\rho, b]$

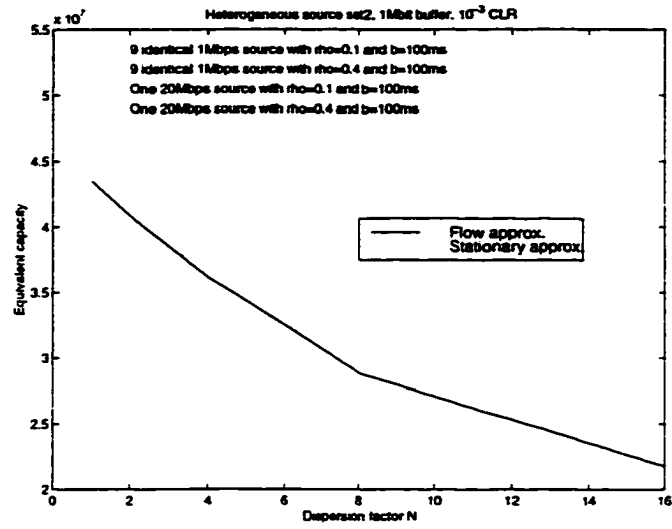


Figure 2.12 Heterogeneous set2 with two different ρ

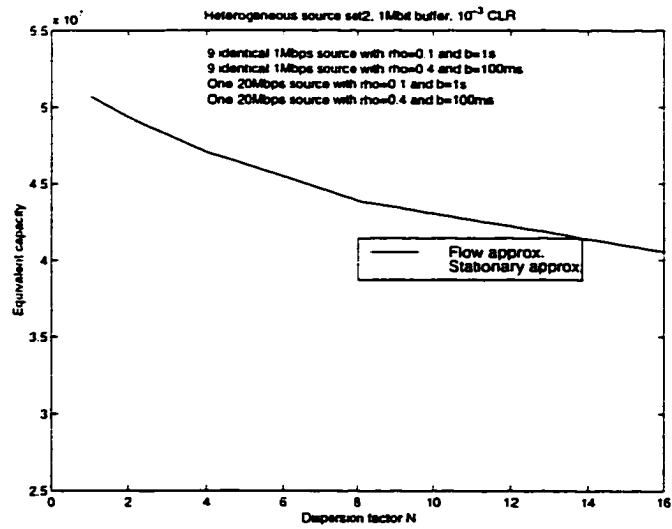


Figure 2.13 Heterogeneous set2 with two different $[\rho, b]$

3 PROBLEM STATEMENT

The focus of this study is on how to reduce CBP (or how to improve bandwidth efficiency for a given CBP requirement) when link capacity is reserved on VP's. Thus, primary objectives of this study are to:

- design efficient routing algorithms including a cost effective traffic dispersion algorithm.
- examine VP capacity reservation and VP network design strategies.
- provide insights of statistical behavior of traffic. This is essential to estimate the accurate equivalent capacity of a VP when both heterogeneous multimedia traffic and traffic dispersion are taken into account.
- explore the changes in traffic characteristics as a result of routing algorithms. When input traffic is dispersed to multiple VP's, the statistical characteristics of traffic seen by each VP is much different from those of input traffic which initially arrived at the system (i.e., the multiplexer). Thus, taking what are given *a priori* as parameters, we want to find quantitative expressions which determine:
 - optimal number of VP's
 - VP capacity required to guarantee the claimed CBP
 - probability of dispersion.
 - statistics of the number of paths taken by a dispersed call
 - statistical characteristics of the traffic seen by each VP as a result of routing

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, our approach to this problem is to 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 (or single-hop routing) algorithms. In direct routing, calls are only allowed to be routed 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 of Fig. 2.2. Successive on-off periods (i.e., $\frac{1}{\beta}$: mean burst length, and $\frac{1}{\alpha}$: mean idle time) are assumed to be mutually independent and identically distributed.
- Call arrivals follow Poisson arrival.
- 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 Guérin *et al.* [1], is used to estimate bandwidth requirement.
- Traffic dispersion is taken into account. 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).
- Followings are assumed to be known *a priori*:
 - a physical ATM network and the capacity of each link
 - statistical distribution and mean value of call arrival rate between a pair of nodes
 - statistical distributions and mean values of input traffic descriptor, $[R_{peak}, \rho, b]$
 - desired cell level QoS (i.e., CLR) and call level QoS (i.e., CBP)
 - parameters of the traffic dispersion algorithm.

4 COST EFFECTIVE DISPERSION ROUTING ALGORITHM

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 which decides the number of dispersion paths upon arriving of a new call. The algorithm determines whether traffic dispersion be 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. 4.1 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:

$$\text{cost} \triangleq \text{Average Free Capacity} \times \text{coefficient} \times (N - 1). \quad (4.1)$$

where coefficient is given as an input parameter of the 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 $\frac{R_{peak}}{N}$. 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 $\frac{R_{peak}}{N}$ 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. In order to calculate the exact traffic dispersion gain, it is assumed that the realistic number of VP's are established for the same class of QoS traffic between a pair of source and destination nodes. Simulation results show that the traffic dispersion gain is not considerable when the number of dispersion paths is larger than 8.

Cost Effective Dispersion (CED)

Assume that $VP_1, VP_2, \dots, VP_i, \dots, VP_M$ are available.

$\Delta = \infty; S = \{\emptyset\};$

$N = 1;$

do

 for each VP_i do

 calculate $\delta_{N,i}$ induced by the new call with $(\frac{R_{peak}}{N}, \rho, b);$

 end

$S_N = \{\emptyset\};$

 select N VP's with the least $\delta_{N,i} \rightarrow S_N;$

$\Delta_N = \sum_{i \in S_N} \delta_{N,i};$

 if $N \geq 2$ then

 if $\Delta_N < \Delta_1 - cost$ and $\Delta_N < \Delta$ then

$\Delta = \Delta_N; S = S_N;$

 endif

 else

$\Delta = \Delta_1; S = S_1;$

 endif

$N = N + 1;$

while $N \leq M$

Figure 4.1 Cost effective dispersion algorithm

5 SIMULATION RESULTS

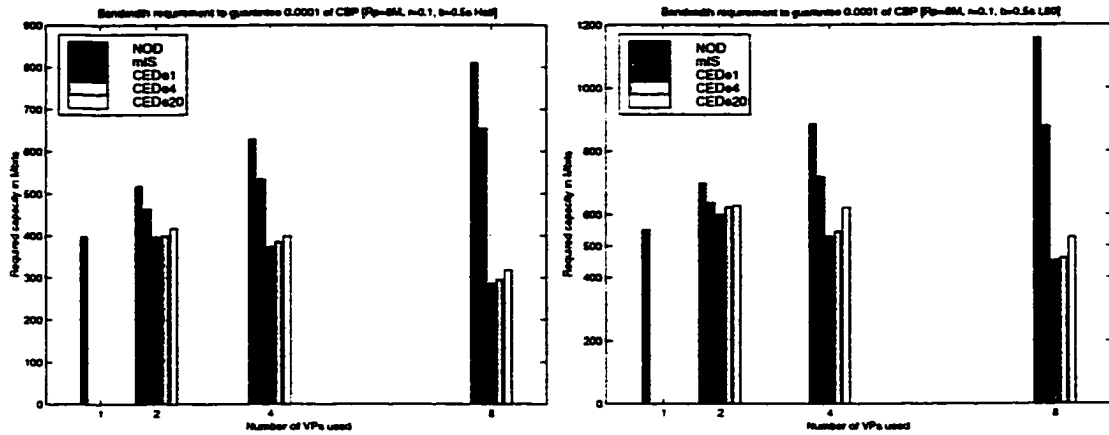
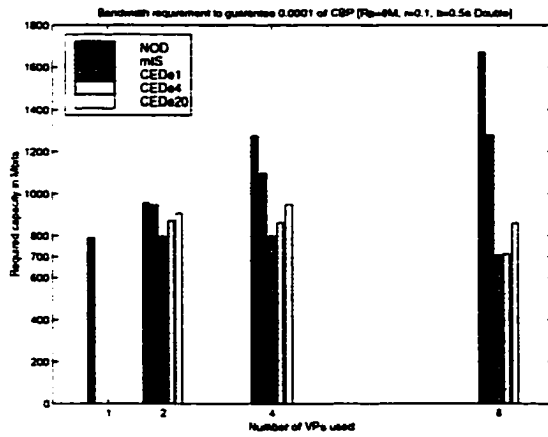
1. Simulation parameters

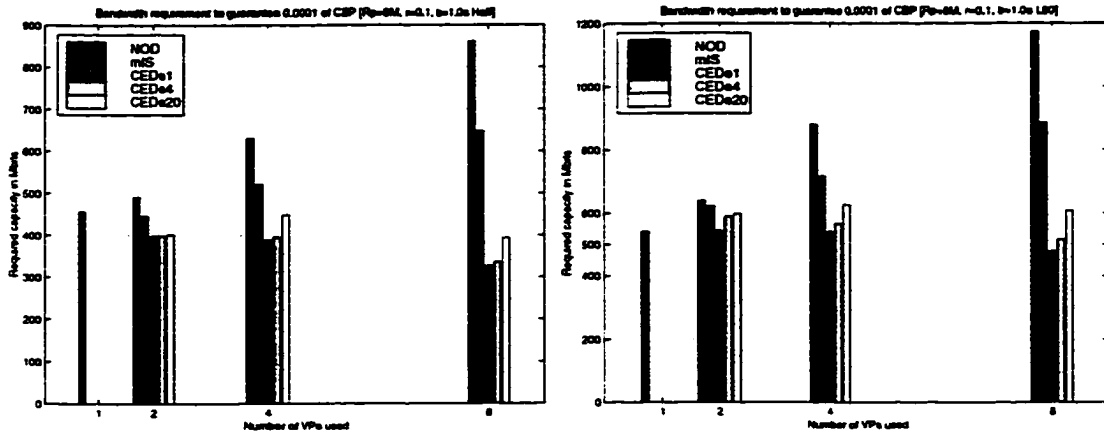
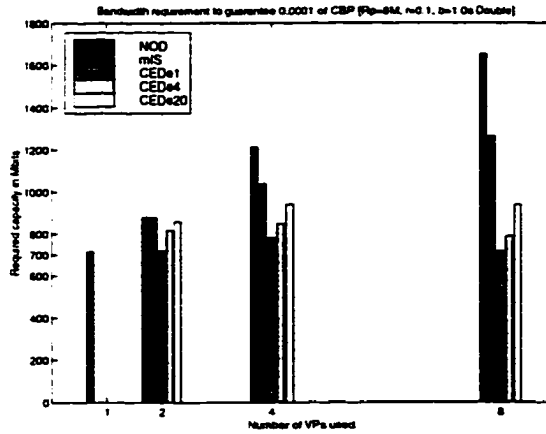
- Routing algorithms
 - For the purpose of comparison, two non-dispersion routing algorithms (i.e., LLP and mIS) are tested. Upon call arrival, LLP selects a least-loaded VP_i while mIS chooses a single path VP_i with the least $\delta_{1,i}$.
 - As a dispersion routing algorithm, CED is used.
- Traffic 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
- Mean call arrival rates were designed for each simulation such that, when LLP is used with 8 VP's, each VP achieves average 80Mbps load. These are reference mean arrival rates. Half and double of these reference mean arrival rates were also used in simulations.
- As desired CBP's, 0.0001, 0.001, and 0.01 were used.

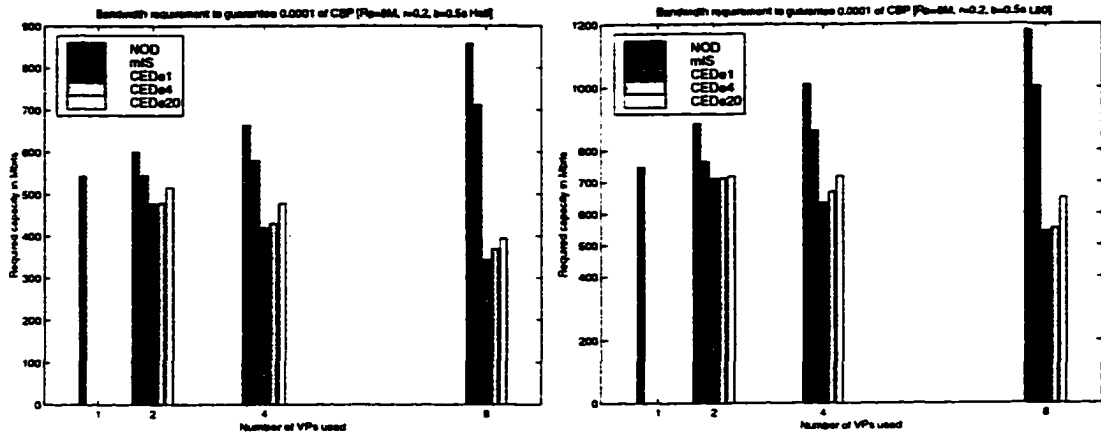
2. Performance of routing algorithms

Simulation results, presented in Fig. 5.1 - Fig. 5.8, show capacity requirements for various traffic 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.

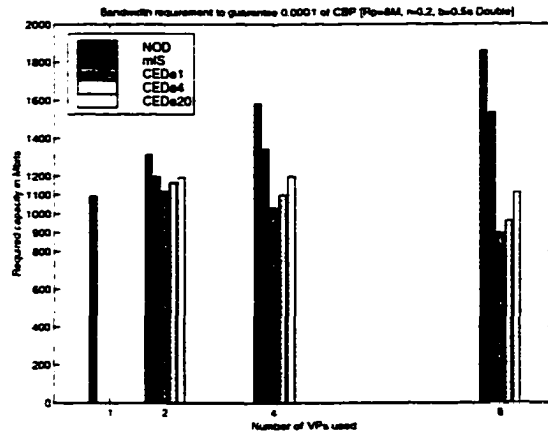
(a) when $\lambda = 1.24$ (b) when $\lambda = 2.48$ (c) when $\lambda = 4.96$ Figure 5.1 Bandwidth requirement when $[R_{peak} = 8Mbps, \bar{\rho} = 0.1, \bar{b} = 0.5sec]$

(a) when $\lambda = 1.08$ (b) when $\lambda = 2.16$ (c) when $\lambda = 4.32$ Figure 5.2 Bandwidth requirement when $[R_{peak} = 8Mbps, \bar{\rho} = 0.1, \bar{b} = 1.0sec]$



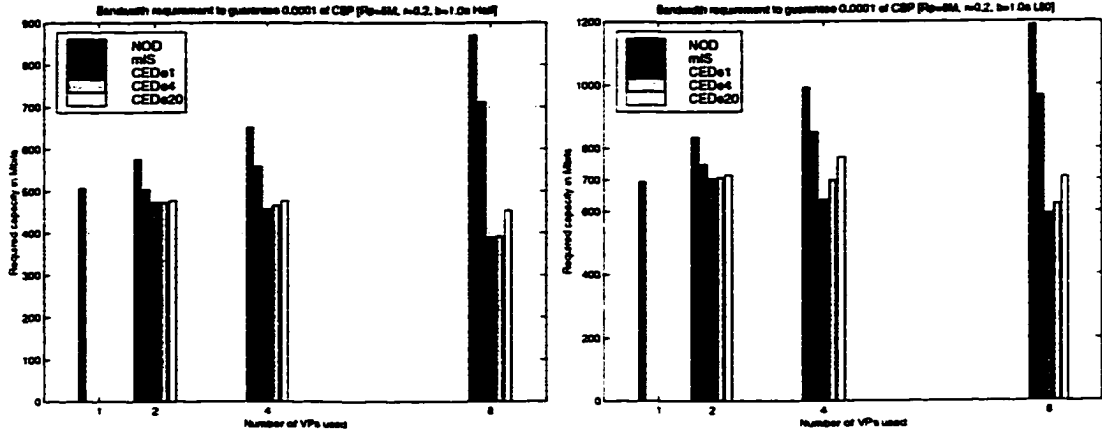
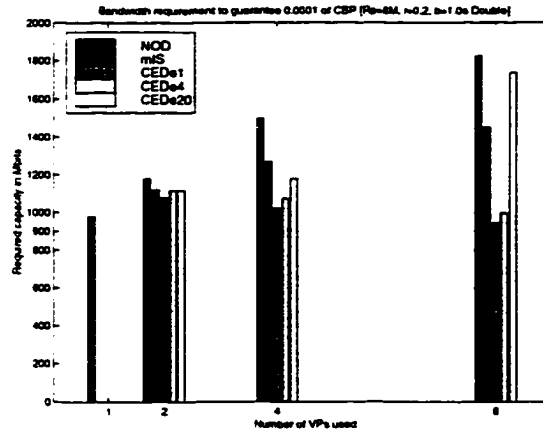
(a) when $\lambda = 1.105$

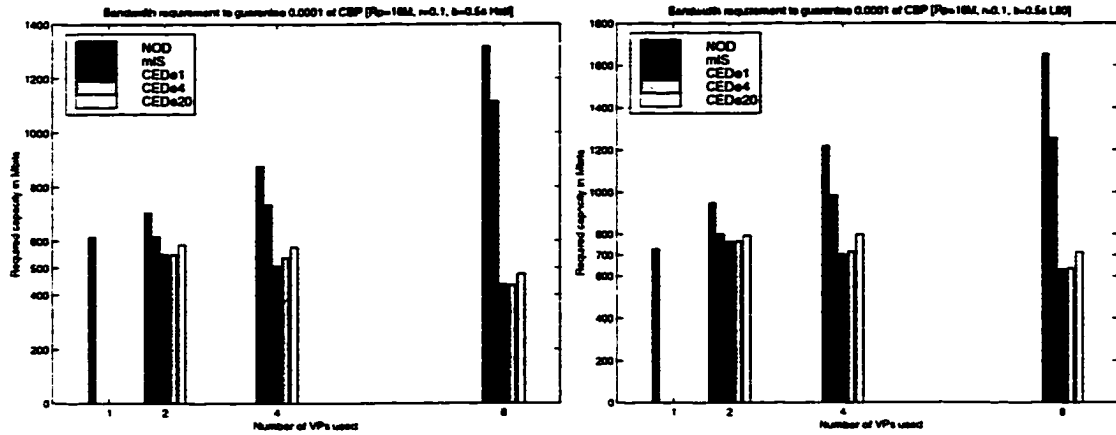
(b) when $\lambda = 2.21$



(c) when $\lambda = 4.42$

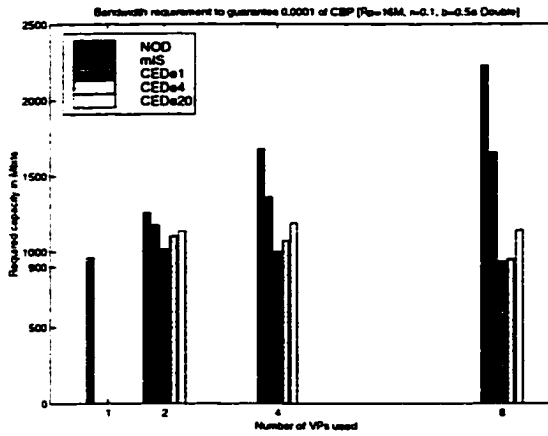
Figure 5.3 Bandwidth requirement when $[R_{peak} = 8Mbps, \bar{\rho} = 0.2, \bar{b} = 0.5sec]$

(a) when $\lambda = 0.94$ (b) when $\lambda = 1.88$ (c) when $\lambda = 3.76$ Figure 5.4 Bandwidth requirement when $[R_{peak} = 8Mbps, \bar{\rho} = 0.2, \bar{b} = 1.0sec]$



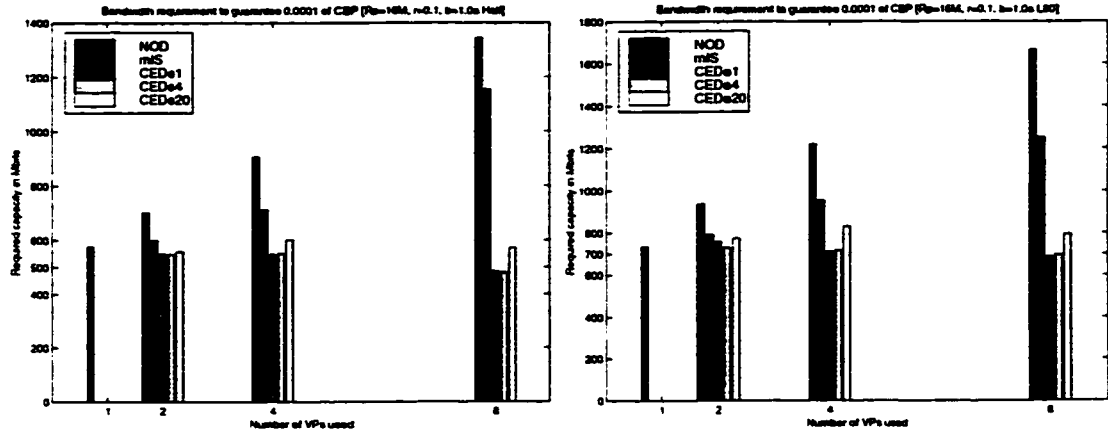
(a) when $\lambda = 0.505$

(b) when $\lambda = 1.01$



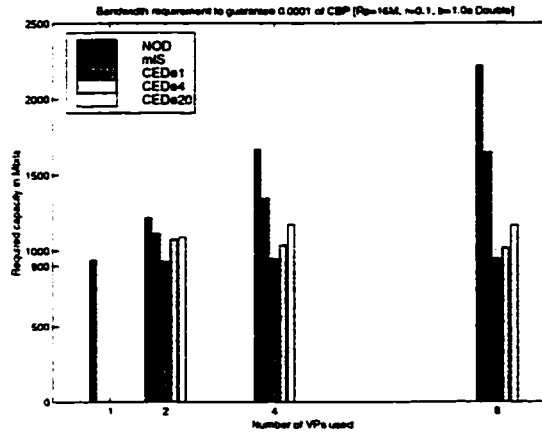
(c) when $\lambda = 2.02$

Figure 5.5 Bandwidth requirement when $[\bar{R}_p = 16Mbps, \bar{\rho} = 0.1, \bar{b} = 0.5sec]$



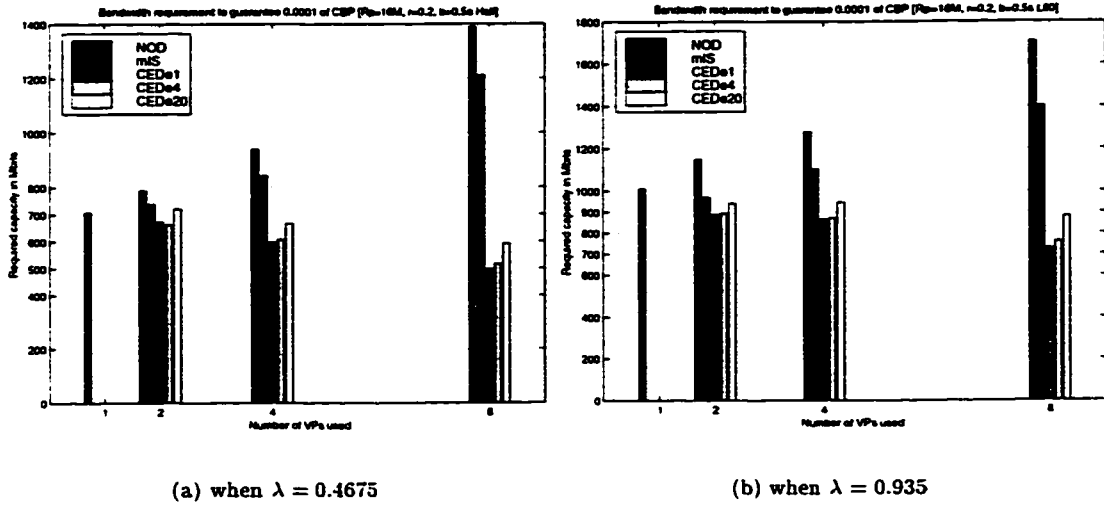
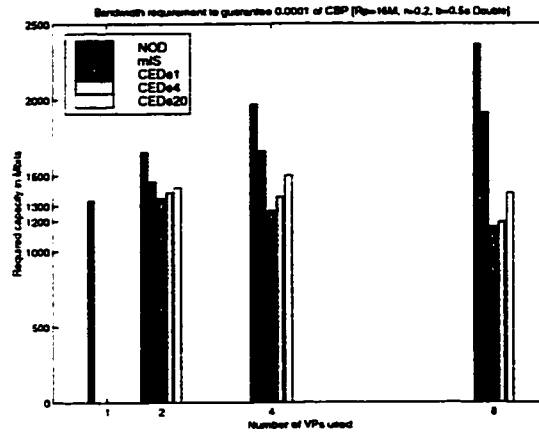
(a) when $\lambda = 0.4525$

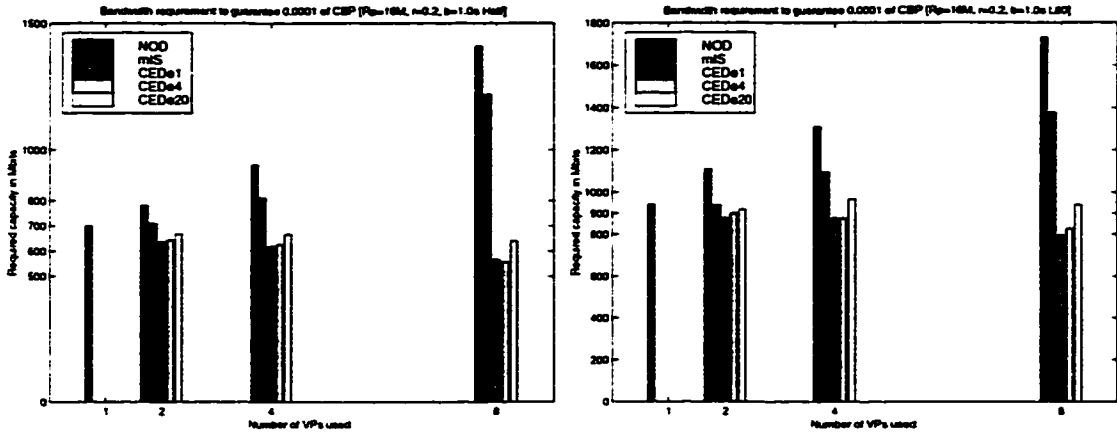
(b) when $\lambda = 0.905$



(c) when $\lambda = 1.81$

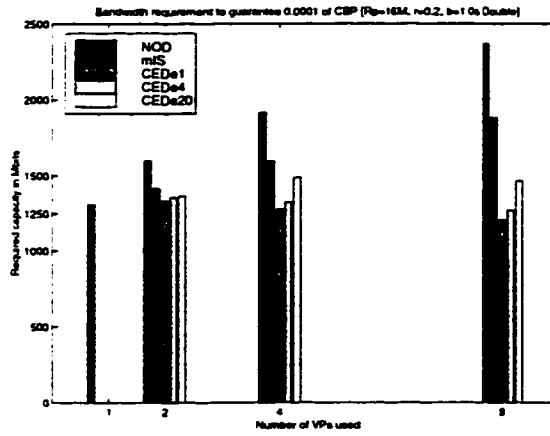
Figure 5.6 Bandwidth requirement when [$\bar{R}_p = 16Mbps, \bar{\rho} = 0.1, \bar{b} = 1.0sec$]

(a) when $\lambda = 0.4675$ (b) when $\lambda = 0.935$ (c) when $\lambda = 1.87$ Figure 5.7 Bandwidth requirement when $[\bar{R}_p = 16Mbps, \bar{\rho} = 0.2, \bar{b} = 0.5sec]$



(a) when $\lambda = 0.4175$

(b) when $\lambda = 0.835$



(c) when $\lambda = 1.67$

Figure 5.8 Bandwidth requirement when $[\bar{R}_p = 16Mbps, \bar{\rho} = 0.2, \bar{b} = 1.0sec]$

- 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 priori* 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.

3. Effect of CED cost coefficient on the dispersion factor D_f

Fig. 5.9 and Fig. 5.10 illustrate the effect of CED coefficient on the dispersion factor D_f , the number of paths taken by a call. These were measured when simulations in Fig. 5.1(a) were performed. In this particular instances, 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 CED cost 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.

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 which 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.

- (a) Fig. 5.11 shows CDF of interarrival time collected at VP_0 when simulations of Fig. 5.1(a) were performed with 8 VP's. Means and standard deviations are summarized at Table 5.1.

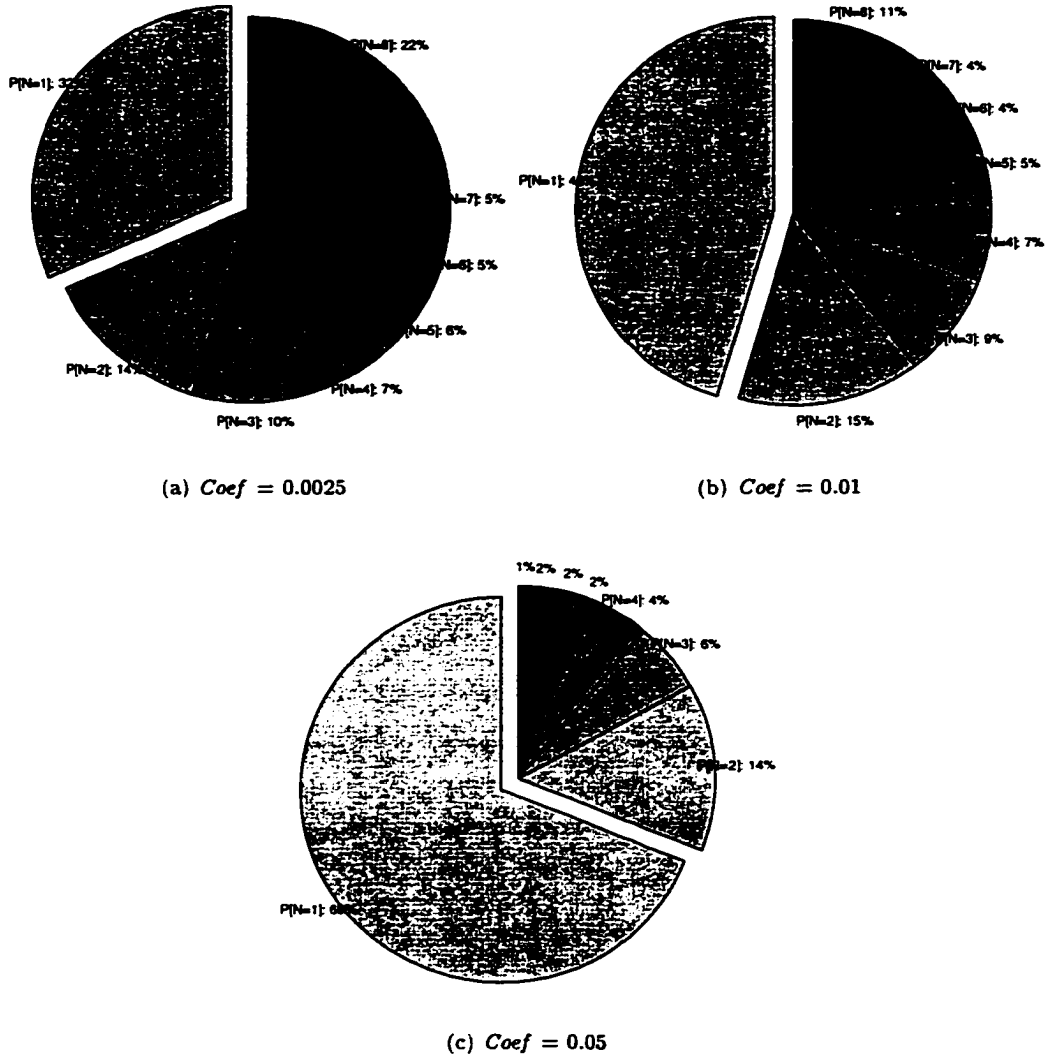


Figure 5.9 Effect of CED coefficient on D_f [8 VP's]

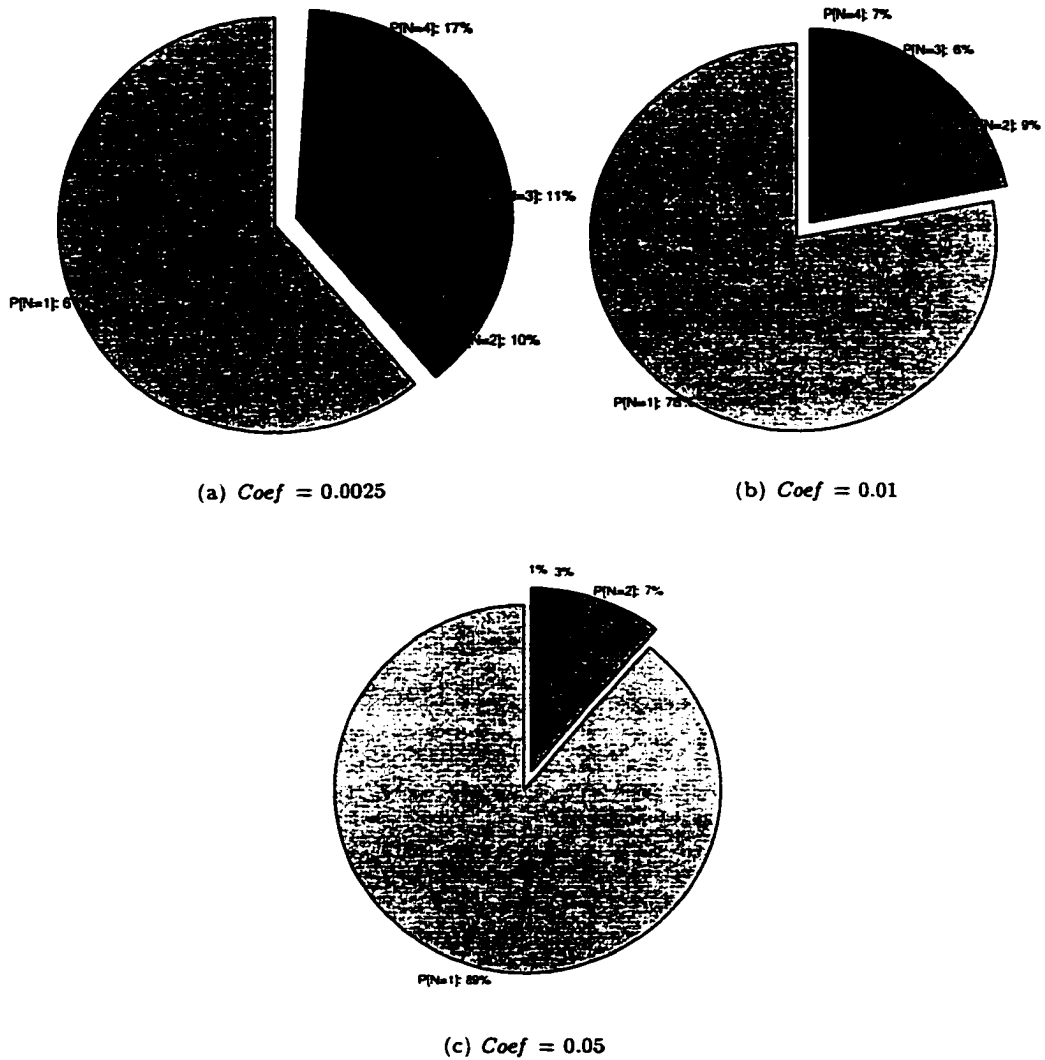
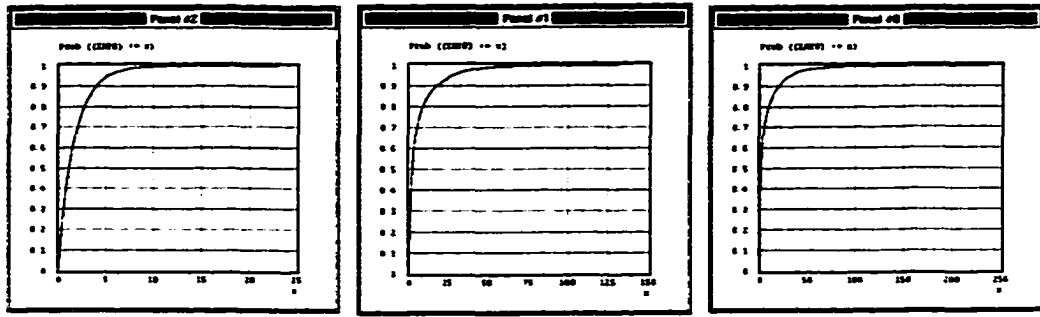


Figure 5.10 Effect of CED coefficient on D_f [4 VP's]

(a) CED with *Coef* 0.0025

(b) mIS

(c) LLP

Figure 5.11 CDF of interarrival time at VP_0

Basically, they are analogous to those of exponential distributions. Note that, before routing, interarrival time of input traffic is exponentially distributed with mean of $\frac{1}{1.24}$. Mean and standard deviation of exponential distributions should be same. When CED is used, these two are about same. From Fig. 5.9(a) 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 $8/(3.86 \times 1.24)$, i.e., 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.

- (b) 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. Fig. 5.12 shows CDF's of R_{peak} collected at VP_0 when simulations of Fig. 5.1(a) were performed with 8 VP's. Means and standard deviations are summarized at Table 5.2. For mIS and LLP, R_{peak} distribution at individual VP was same as that of input traffic.
- (c) The distributions of source utilization ρ (Fig. 5.13 and Table 5.3) and mean burst period b (Fig. 5.14 and Table 5.4) 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 (Table 5.5).

Table 5.1 Mean (σ) of interarrival time at VP_0

| | VP0 | VP1 | VP2 | VP3 | VP4 | VP5 | VP6 | VP7 |
|------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| CED | 1.67 (1.79) | 1.66 (1.76) | 1.66 (1.75) | 1.66 (1.75) | 1.67 (1.77) | 1.67 (1.77) | 1.66 (1.76) | 1.67 (1.77) |
| mIS | 6.40 (10.54) | 6.91 (11.49) | 8.03 (12.64) | 9.69 (11.80) | 7.45 (11.61) | 6.84 (11.46) | 5.18 (9.37) | 4.97 (9.18) |
| LLP | 6.61 (13.39) | 6.25 (12.08) | 6.53 (13.39) | 6.54 (12.92) | 6.36 (12.15) | 6.51 (12.74) | 6.40 (12.86) | 6.65 (12.81) |

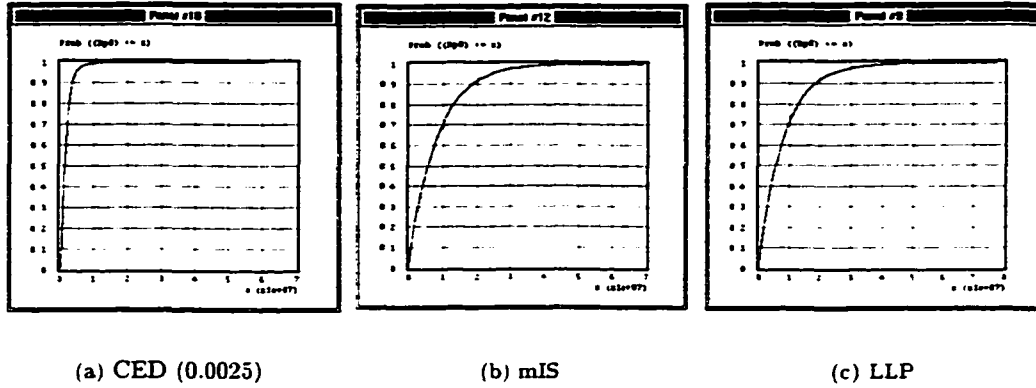
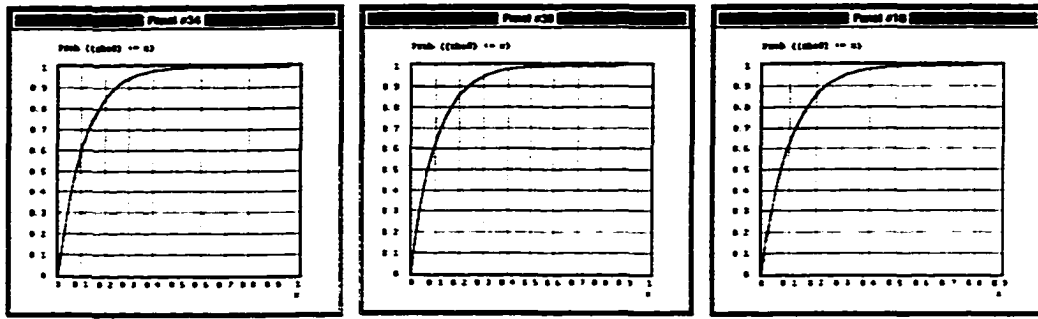


Figure 5.12 CDF of R_{peak} at VP_0

Table 5.2 Mean (σ) of R_{peak} at VP_0

| | VP0 | VP1 | VP2 | VP3 | VP4 | VP5 | VP6 | VP7 |
|------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| CED | 2.08 (1.72) | 2.07 (1.72) | 2.07 (1.64) | 2.07 (1.62) | 2.07 (1.67) | 2.07 (1.65) | 2.07 (1.68) | 2.06 (1.64) |
| mIS | 8.05 (7.96) | 8.10 (8.11) | 8.32 (8.46) | 8.23 (8.35) | 8.21 (8.26) | 8.12 (8.15) | 7.82 (7.71) | 7.68 (7.44) |
| LLP | 8.11 (8.19) | 7.93 (7.99) | 7.91 (8.03) | 8.00 (8.00) | 7.95 (7.94) | 7.96 (7.89) | 7.93 (7.95) | 8.12 (8.00) |



(a) CED (0.0025)

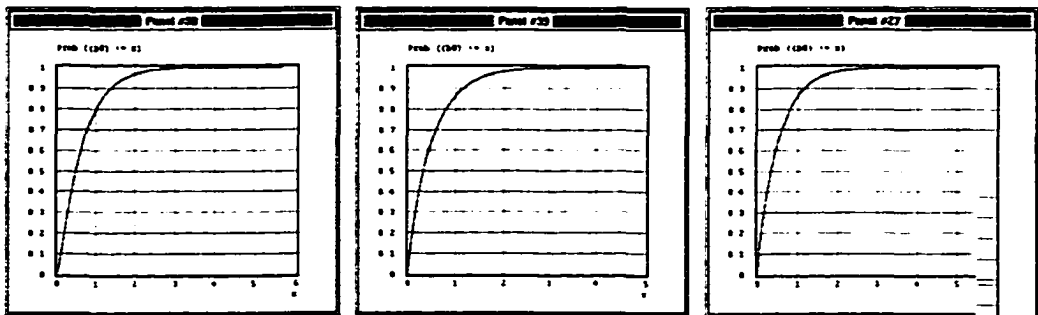
(b) mIS

(c) LLP

Figure 5.13 CDF of ρ at VP_0

Table 5.3 Mean (σ) of ρ at VP_0

| | VP0 | VP1 | VP2 | VP3 | VP4 | VP5 | VP6 | VP7 |
|-----|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| CED | 0.11 (0.10) | 0.11 (0.10) | 0.11 (0.10) | 0.11 (0.10) | 0.11 (0.10) | 0.11 (0.10) | 0.11 (0.10) | 0.11 (0.10) |
| mIS | 0.10 (0.10) | 0.10 (0.10) | 0.11 (0.10) | 0.11 (0.10) | 0.11 (0.11) | 0.10 (0.10) | 0.09 (0.09) | 0.09 (0.09) |
| LLP | 0.10 (0.10) | 0.10 (0.10) | 0.10 (0.10) | 0.10 (0.10) | 0.10 (0.10) | 0.10 (0.10) | 0.10 (0.10) | 0.10 (0.10) |



(a) CED (0.0025)

(b) mIS

(c) LLP

Figure 5.14 CDF of b at VP_0

Table 5.4 Mean (σ) of b at VP_0

| | VP0 | VP1 | VP2 | VP3 | VP4 | VP5 | VP6 | VP7 |
|------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| CED | 0.65 (0.55) | 0.65 (0.55) | 0.64 (0.55) | 0.64 (0.54) | 0.64 (0.55) | 0.64 (0.54) | 0.64 (0.55) | 0.64 (0.55) |
| mIS | 0.49 (0.50) | 0.50 (0.50) | 0.42 (0.46) | 0.44 (0.47) | 0.43 (0.47) | 0.48 (0.49) | 0.57 (0.52) | 0.60 (0.51) |
| LLP | 0.50 (0.51) | 0.50 (0.50) | 0.50 (0.49) | 0.50 (0.50) | 0.50 (0.50) | 0.50 (0.50) | 0.50 (0.50) | 0.51 (0.51) |

Table 5.5 Mean (σ) of b at VP_0 with 4 VP's

| | VP0 | VP1 | VP2 | VP3 |
|------------|-----------------------|-----------------------|-----------------------|-----------------------|
| CED | 0.60 (0.50) | 0.48 (0.45) | 0.41 (0.40) | 0.46 (0.43) |
| mIS | 0.53 (0.50) | 0.62 (0.52) | 0.30 (0.40) | 0.40 (0.47) |
| LLP | 0.50 (0.50) | 0.50 (0.50) | 0.51 (0.50) | 0.50 (0.50) |

6 ANALYTICAL MODELS

The simulation study, presented in section 5, demonstrates how traffic dispersion routing algorithms improve bandwidth efficiency significantly for a given CBP requirement (or reduce CBP for a given capacity reserved for the VP), when establishing more than one VP's for each different class of QoS is possible. With the efficient traffic dispersion routing algorithm., such as CED proposed in section 4, network engineers will look for the cost effective VP network design strategies. In order to sketch the required capacity and the optimal number of VP's, an accurate and tractable analytical model must be developed.

Providing insights of statistical behavior of traffic is the primary objective of developing the analytical model. This is essential to estimate the accurate equivalent capacity of a VP when both heterogeneous multimedia traffic and traffic dispersion are taken into account. Furthermore, we would like to explore the effect of routing algorithm on the characteristics of input traffic. When input traffic is dispersed to multiple VP's, the statistical characteristics of the traffic seen by each VP is much different from those of input traffic arrived at the multiplexer. If round-robin routing is used and traffic dispersion is performed with fixed dispersion factor for every arriving call, it would be simple to predict the statistical characteristics of traffic stream for each VP, However, for any dynamic routing algorithm, it poses a formidable challenge.

Followings are assumed to be known *a priori*:

- a physical ATM network and the capacity of each link
- statistical distribution and mean value of call arrival rate between each pair of nodes
- statistical distributions and mean values of input traffic descriptor, $[R, \rho, b]$
- desired cell level QoS (i.e., CLR) and call level QoS (i.e., CBP)
- parameters of the traffic dispersion algorithm.

Taking what are given above as parameters, we want to find quantitative expressions which determine:

- optimal number of VP's
- VP capacity required to guarantee the claimed CBP
- probability of dispersion.
- statistics of the number of paths taken by each call
- statistical characteristics of the traffic seen by each VP. such as
 - distribution of call arrival rate
 - distribution of peak rate, R
 - distribution of mean source utilization, ρ
 - distribution of mean burst period, b .

If we consider only homogeneous traffic and fluid-flow approximation as in [5. 9. 10]. a VP of an one-stage multiplexer, as illustrated in Fig. 2.7(b), can be represented as a continuous-time Markov chain shown in Fig. 6.1. Then analysis would be straightforward. This is because each state can be represented by a single parameter. the number of connections. The equivalent capacity of each state is simply obtained by multiplying the equivalent capacity of a connection by the number connections at the state. With assumptions of homogeneous traffic and fluid-flow approximation, each connection requires an identical equivalent capacity which is estimated by Eq.(2.2). Thus, upon call arrival, routing (i.e., determining p_{ij} 's in Fig. 6.1) is performed with only knowledge of the number of connections in each VP. State transitions are determined by λp_{ij} and the reciprocal of call holding time (i.e., service rate μ), where λ is a given mean Poisson arrival rate and call holding times are exponentially distributed with mean $\frac{1}{\mu}$. Once we solve this queueing system in terms of stationary state probability, π_{ij} , required VP capacity for a given CBP are easily determined by investigating the Q-function of π .

However, for heterogeneous traffic, the number of connections does not give any deterministic information to estimate the equivalent capacity. It is estimated by Eq.(2.1), and its calculation requires the distributions of peak rate, R , mean source utilization, ρ , and mean burst period, b .

One may consider a Markov chain representation. But, even for the moderately accurate model, explosion of the number of states and extremely complicated state transitions make this approach intractable. Yet there is no analytical model reported in the literature, which can describe the statistical behavior of heterogeneous traffic as a result of routing algorithms with traffic dispersion. Traditional teletraffic analysis methodologies are unsuitable for modeling this problem.

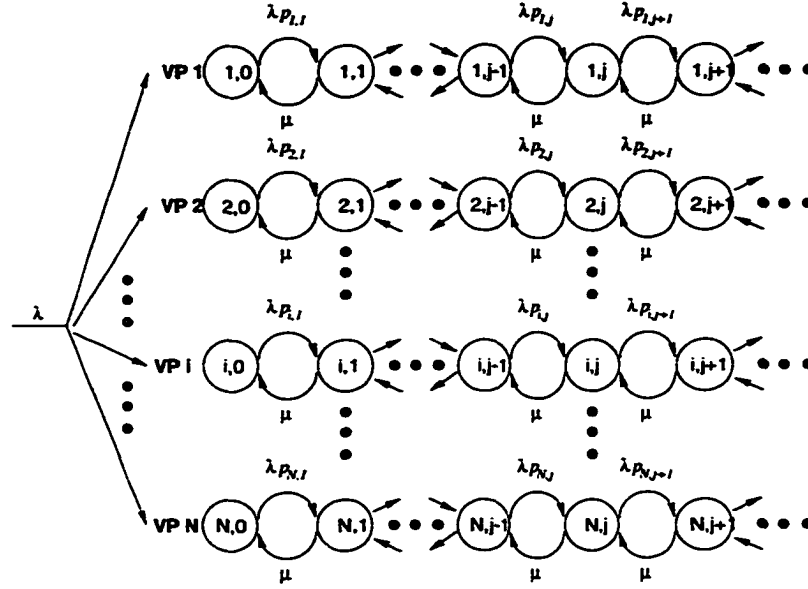


Figure 6.1 Markov chain for a multiplexer with N VP's

Consequently, one must solve the functions of several random variables. Since a function of random variables is also a random variable, if we let traffic descriptor $[R, \rho, b]$ be a three-dimensional random variable, then \hat{C} in Eq.(2.1), $\hat{C}_{(F)}$ in Eq.(2.2), and $\hat{C}_{(S)}$ in Eq.(2.3) are also random variables. Then, it is obvious that, if we obtain Cumulative Density Function (CDF) of the load distribution (i.e., CDF of \hat{C}) of a VP, VP capacity required to meet a given call level QoS (i.e., CBP) is determined readily by investigating the point where the CDF becomes greater than $1 - CBP$.

Single path load distribution

In order to illustrate the method, we assume, for the time being, that R , ρ , and b are exponentially distributed random variables with mean values of \bar{R} , $\bar{\rho}$, and \bar{b} , respectively. This is when the traffic dispersion is not used. If CED is used, R is no longer an exponentially distributed random variable. We will investigate this shortly later. Furthermore, we assume that the equivalent capacity of the VP is largely determined by stationary approximation, $\hat{C}_{(S)}$.

Eq.(2.3) can be represented equivalently as:

$$\hat{C}_{(S)N} = \sum_{k=1}^N R_k \rho_k + K \sqrt{\sum_{k=1}^N R_k^2 \rho_k (1 - \rho_k)}, \quad (6.1)$$

where N is a *i.i.d.* random variable.

If we define random variables \mathcal{M}_N and \mathcal{S}_N as:

$$\begin{aligned}\mathcal{M}_N &= \sum_{k=1}^N R_k \rho_k, \\ \mathcal{S}_N &= \sqrt{\sum_{k=1}^N R_k^2 \rho_k (1 - \rho_k)}.\end{aligned}\tag{6.2}$$

then,

$$\hat{\mathcal{C}}_{(S)N} = \mathcal{M}_N + K \mathcal{S}_N.\tag{6.3}$$

Now, $\hat{\mathcal{C}}_{(S)N}$ becomes the sum of two correlated random variables \mathcal{M}_N and \mathcal{S}_N . Thus, the pdf of $\hat{\mathcal{C}}_{(S)N}$ can be obtained by the linear transformation method, if we find the proper joint pdf of \mathcal{M}_N and \mathcal{S}_N as well as the marginal pdf's of them. In the following, we will see how these are obtained.

1. Probability density function of \mathcal{M}_N

Random variable \mathcal{M}_N is a function of two independent random variables \mathcal{R} and ρ . In detail, it is the sum of a random number of *i.i.d.* random variable:

$$\mathcal{X} = \mathcal{R}\rho.\tag{6.4}$$

(a) pdf of $\mathcal{R}\rho$

In order to find the pdf of \mathcal{M}_N , we need to get the pdf of \mathcal{X} first. Assume $\rho = \varrho$, then $\mathcal{X} = \mathcal{R}\varrho$ is simply a scaled version of \mathcal{R} . The event $\{\mathcal{X} \leq x\}$ occurs when $A = \{\mathcal{R}\varrho \leq x\}$ occurs. Thus, the CDF of \mathcal{X} is given by:

$$\begin{aligned}F_{\mathcal{X}}(x|\varrho) &= P\left[\mathcal{R} \leq \frac{x}{\varrho}\right] \\ &= F_{\mathcal{R}}\left(\frac{x}{\varrho}\right).\end{aligned}\tag{6.5}$$

We can find the pdf of \mathcal{X} by differentiating it with respect to x :

$$f_{\mathcal{X}}(x|\varrho) = \frac{1}{\varrho} f_{\mathcal{R}}\left(\frac{x}{\varrho}\right).\tag{6.6}$$

The pdf of \mathcal{X} is therefore:

$$\begin{aligned} f_{\mathcal{X}}(x) &= \int_{-\infty}^{\infty} \frac{1}{\varrho} f_{\mathcal{R}}\left(\frac{x}{\varrho}\right) f_{\rho}(\varrho) d\varrho \\ &= \int_{-\infty}^{\infty} \frac{1}{\varrho} f_{\mathcal{R},\rho}\left(\frac{x}{\varrho}, \varrho\right) d\varrho. \end{aligned} \quad (6.7)$$

We now use that fact that \mathcal{R} and ρ are independent and exponentially distributed with means $\tilde{\mathcal{R}}$ and $\bar{\rho}$, respectively. Also, we know that $0 < \mathcal{R} < \infty$, and $0 < \rho \leq 1$. Thus, the pdf of $\mathcal{R}\rho$ is:

$$f_{\mathcal{X}}(x) = \int_{0+}^1 \frac{1}{\varrho} f_{\mathcal{R}}\left(\frac{x}{\varrho}\right) f_{\rho}(\varrho) d\varrho \quad x > 0. \quad (6.8)$$

If we expand pdf's of \mathcal{R} and ρ :

$$f_{\mathcal{X}}(x) = \frac{1}{\tilde{\mathcal{R}}\bar{\rho}} \int_{0+}^1 \frac{1}{\varrho} e^{-(\varrho/\bar{\rho} + x/(\tilde{\mathcal{R}}\varrho))} d\varrho. \quad (6.9)$$

Since this integration would not provide closed form solution, numerical solution is used in this study.

(b) pdf of $\mathcal{M}_{\mathcal{N}}$

From Eq.(5.14) in [35], we know that the characteristic function of the sum of N random numbers of \mathcal{X} is found by evaluating the generating function of N at $z = \Phi_{\mathcal{X}}(\omega)$. The characteristic function of \mathcal{X} , $\Phi_{\mathcal{X}}(\omega)$ is:

$$\begin{aligned} \Phi_{\mathcal{X}}(\omega) &= E[e^{j\omega x}] \\ &= \int_{-\infty}^{\infty} f_{\mathcal{X}}(x) e^{j\omega x} dx. \end{aligned} \quad (6.10)$$

where $f_{\mathcal{X}}(x)$ is the pdf of random variable \mathcal{X} .

If we assume that the call arrival process of a VP is a Poisson process with mean λ , the number of connections at a VP can be modeled as M/M/ ∞ queue, also called delay center. When mean service time (i.e., exponentially distributed call holding time) is \bar{s} , traffic intensity τ is $\lambda\bar{s}$ by definition. Then stationary state probability π_n (i.e., probability of n connections in the VP) is:

$$\pi_n = \frac{\tau^n e^{-\tau}}{n!}. \quad (6.11)$$

This is the familiar probability mass function (pmf) of Poisson random variable with mean τ , and its probability generating function is given by:

$$G_N(z) = e^{\tau(z-1)} \quad (6.12)$$

Therefore, the characteristic function of $\mathcal{M}_N = \sum_{k=1}^N X_k$ can be found by evaluating Eq.(6.12) at $z = \Phi_{\mathcal{X}}(\omega)$:

$$\begin{aligned} \Phi_{\mathcal{M}_N}(\omega) &= G_N(\Phi_{\mathcal{X}}(\omega)) \\ &= e^{\lambda \bar{s}(\Phi_{\mathcal{X}}(\omega)-1)}. \end{aligned} \quad (6.13)$$

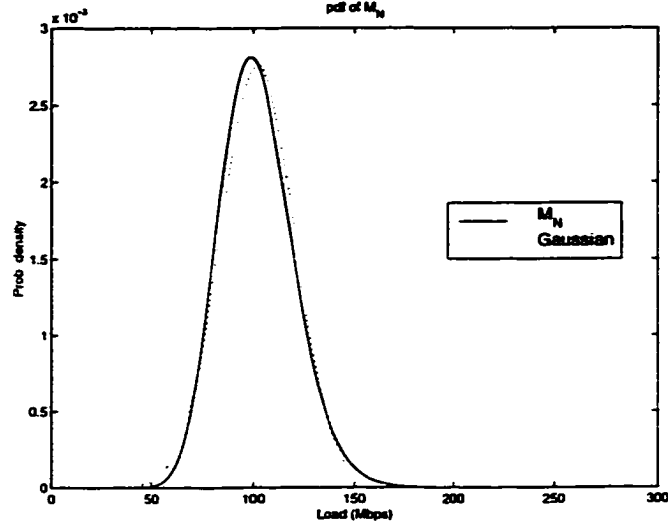
Now we can obtain the pdf of \mathcal{M}_N by taking the inverse transform of $\Phi_{\mathcal{M}_N}(\omega)$:

$$f_{\mathcal{M}_N}(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \Phi_{\mathcal{M}_N}(\omega) e^{-j\omega x} d\omega. \quad (6.14)$$

Fig. 6.2 shows the pdf of \mathcal{M}_N when 8 Mbps, 0.1, 0.5 sec., and 2.48 are used for \bar{R} , $\bar{\rho}$, \bar{b} , and λ , respectively. In the figure, dotted line depicts the Gaussian pdf with mean and variance of \mathcal{M}_N . It is not amazing that the pdf of \mathcal{M}_N is approximately the Gaussian pdf. As defined by Eq.(6.2), \mathcal{M}_N is the sum of N random variable \mathcal{X} . With 60 sec. mean service time \bar{s} , the average number of connection is about 150, which is large enough to apply the *Central Limit Theorem*. The theorem tells, as long as \mathcal{X} has a finite mean and finite variance, the CDF of a properly normalized \mathcal{M}_N approaches that of a Gaussian random variable. It is obvious that the random variable \mathcal{X} , as defined by Eq.(6.4), has finite mean and variance.

2. Probability density function of $K\mathcal{S}_N$

To get the pdf of $K\mathcal{S}_N$, we need to find the pdf of $\mathcal{R}^2\rho(1-\rho)$. Let's define random variables \mathcal{U} , \mathcal{V} , and \mathcal{W} as follows:

Figure 6.2 Distribution of M_N

$$\begin{aligned}
 \mathcal{W} &= \mathcal{R}^2, \\
 \mathcal{U} &= \rho(1 - \rho), \\
 \mathcal{V} &= \sigma^2 \\
 &= \mathcal{R}^2 \rho(1 - \rho) \\
 &= \mathcal{W}\mathcal{U}.
 \end{aligned} \tag{6.15}$$

(a) pdf of $\mathcal{W} = \mathcal{R}^2$

The event $\{\mathcal{W} \leq w\}$ occurs when $\{\mathcal{R}^2 \leq w\}$ or equivalently when $\{-\sqrt{w} \leq \mathcal{R} \leq \sqrt{w}\}$ for w nonnegative. The event is null when w is negative. Thus the CDF of \mathcal{W} is:

$$\begin{aligned}
 F_{\mathcal{W}}(w) &= 0 & \text{for } w < 0 \\
 &= F_{\mathcal{R}}(\sqrt{w}) - F_{\mathcal{R}}(-\sqrt{w}) & \text{for } w > 0.
 \end{aligned} \tag{6.16}$$

We get the pdf of \mathcal{W} by differentiating it with respect to w :

$$f_{\mathcal{W}}(w) = \frac{f_{\mathcal{R}}(\sqrt{w})}{2\sqrt{w}} \quad \text{for } w > 0. \tag{6.17}$$

We used the fact that $0 < \mathcal{R} < \infty$. Since \mathcal{R} is an exponentially distributed random variable with mean $\bar{\mathcal{R}}$, Eq.(6.17) turns into:

$$f_{\mathcal{W}}(w) = \frac{1}{2\bar{\mathcal{R}}\sqrt{w}} e^{-\sqrt{w}/\bar{\mathcal{R}}}. \quad (6.18)$$

(b) pdf of $\mathcal{U} = \rho(1 - \rho)$

If we investigate the relationship between \mathcal{U} and ρ , we can get the CDF of \mathcal{U} :

$$\begin{aligned} F_{\mathcal{U}}(u) &= 0 & \text{for } u < 0 \\ & 1 & \text{for } u \geq 0 \\ & 1 - F_{\rho}\left(\frac{1 + \sqrt{1 - 4u}}{2}\right) + F_{\rho}\left(\frac{1 - \sqrt{1 - 4u}}{2}\right). \end{aligned} \quad (6.19)$$

And if we differentiate it with respect to u , we obtain the pdf of \mathcal{U} :

$$\begin{aligned} f_{\mathcal{U}}(u) &= \frac{1}{\sqrt{1 - 4u}} \left\{ f_{\rho}\left(\frac{1 + \sqrt{1 - 4u}}{2}\right) + f_{\rho}\left(\frac{1 - \sqrt{1 - 4u}}{2}\right) \right\} \\ &= \frac{1}{\bar{\rho}\sqrt{1 - 4u}} \left\{ e^{-\frac{1}{2\bar{\rho}}(1 + \sqrt{1 - 4u})} + e^{-\frac{1}{2\bar{\rho}}(1 - \sqrt{1 - 4u})} \right\}. \end{aligned} \quad (6.20)$$

(c) pdf of $\mathcal{V} = \mathcal{R}^2\rho(1 - \rho)$

As shown above, we found the marginal pdf's of \mathcal{W} and \mathcal{U} . Since $\mathcal{V} = \mathcal{W}\mathcal{U}$, as well as \mathcal{W} and \mathcal{U} are independent, if we follow the same approach that is used to get Eq.(6.9), we can find the pdf of \mathcal{V} as:

$$\begin{aligned} f_{\mathcal{V}}(v) &= \int_{0^+}^{\frac{1}{v}} \frac{1}{u} f_{\mathcal{W}}\left(\frac{v}{u}\right) f_{\mathcal{U}}(u) du \\ &= \frac{1}{2\bar{\mathcal{R}}\bar{\rho}} \int_{0^+}^{\frac{1}{v}} \frac{1}{\sqrt{vu(1 - 4u)}} e^{-\frac{1}{\bar{\rho}}\sqrt{\frac{v}{u}}} \left\{ e^{-\frac{1}{2\bar{\rho}}(1 + \sqrt{1 - 4u})} + e^{-\frac{1}{2\bar{\rho}}(1 - \sqrt{1 - 4u})} \right\} du. \end{aligned} \quad (6.21)$$

(d) pdf of $\mathcal{V}_{\mathcal{N}} = \sum_{k=1}^{\mathcal{N}} R_k^2 \rho_k (1 - \rho_k)$

The characteristic function of $\mathcal{V}_{\mathcal{N}}$ can be found by evaluating the generating function of \mathcal{N} Eq.(6.12) at $z = \Phi_{\mathcal{V}}(\omega)$:

$$\begin{aligned} \Phi_{\mathcal{V}_{\mathcal{N}}}(\omega) &= G_{\mathcal{N}}(\Phi_{\mathcal{V}}(\omega)) \\ &= e^{\lambda \bar{z} (\Phi_{\mathcal{V}}(\omega) - 1)}. \end{aligned} \quad (6.22)$$

The pdf of $\mathcal{M}_{\mathcal{N}}$ is obtained by taking the inverse transform of $\Phi_{\mathcal{V}_{\mathcal{N}}}(\omega)$:

$$f_{\mathcal{V}_N}(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \Phi_{\mathcal{V}_N}(\omega) e^{-j\omega x} d\omega. \quad (6.23)$$

(e) pdf of $K\mathcal{S}_N = K\sqrt{\mathcal{V}_N}$

If we investigate the relationship between $K\mathcal{S}_N$ and \mathcal{V}_N , we can find that the CDF of $K\mathcal{S}_N$ is related to that of \mathcal{V}_N as follows:

$$F_{K\mathcal{S}_N}(x) = F_{\mathcal{V}_N}\left(\frac{x^2}{K^2}\right). \quad (6.24)$$

Thus, the pdf of $K\mathcal{S}_N$ is:

$$f_{K\mathcal{S}_N}(x) = \frac{2x}{K^2} f_{\mathcal{V}_N}\left(\frac{x^2}{K^2}\right). \quad (6.25)$$

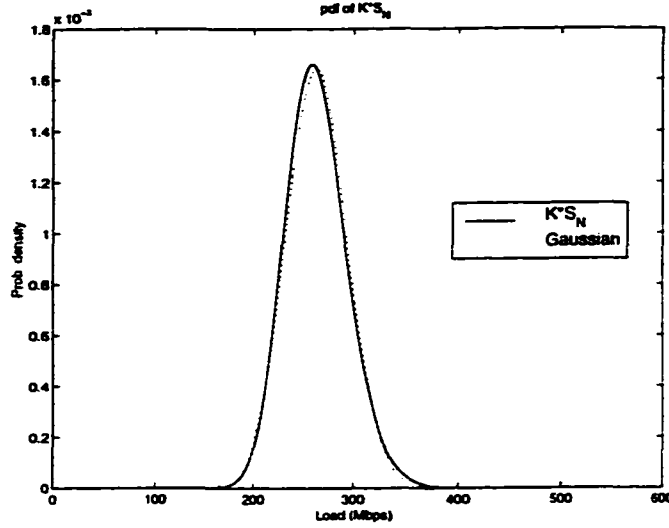
Fig. 6.3 shows the pdf of $K\mathcal{S}_N$ when 8 Mbps, 0.1, 0.5 sec., and 2.48 are used for \bar{R} , $\bar{\rho}$, \bar{b} , and λ , respectively. In the figure, dotted line depicts the Gaussian pdf with mean and variance of $K\mathcal{S}_N$. By same reason, the pdf of $K\mathcal{S}_N$ is approximately the Gaussian pdf.

3. Probability density function of $\hat{\mathcal{C}}_{(S),N}$

So far, we obtained pdf's of \mathcal{M}_N and $K\mathcal{S}_N$. Now we want to get the pdf of $\hat{\mathcal{C}}_{(S),N}$. As defined early in this section, $\hat{\mathcal{C}}_{(S),N}$ is the sum of two random variables \mathcal{M}_N and $K\mathcal{S}_N$. And the pdf of it can be found by linear transform method. Before using the transform method, however, we have to know the joint pdf of \mathcal{M}_N and $K\mathcal{S}_N$. In general, the joint pdf cannot be obtained from marginal pdf's. We know, as shown in Fig. 6.2 and Fig. 6.3 above, that \mathcal{M}_N and $K\mathcal{S}_N$ are approximately Gaussian random variables. From simulation, we also learned that link load follows Gaussian distribution. Furthermore, it is well known that the sum of jointly Gaussian random variables is always a Gaussian random variable. For example, if $X_1, X_2, \dots, X_i, \dots, X_n$ are jointly Gaussian random variables, $Z = a_1 X_1 + a_2 X_2 + \dots + a_i X_i + \dots + a_n X_n$ is a Gaussian random variable. If one noticed these facts, it is not difficult to predict that \mathcal{M}_N and $K\mathcal{S}_N$ are most likely jointly Gaussian random variables.

The random variables $X_1, X_2, \dots, X_i, \dots, X_n$ are said to be jointly Gaussian, if their joint pdf is given by:

$$f_{\mathbf{X}}(\mathbf{x}) \triangleq f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n) = \frac{\exp\{-\frac{1}{2}(\mathbf{x}-\mathbf{m})^T \boldsymbol{\kappa}^{-1}(\mathbf{x}-\mathbf{m})\}}{(2\pi)^{\frac{n}{2}} |\boldsymbol{\kappa}|^{\frac{1}{2}}}, \quad (6.26)$$

Figure 6.3 Distribution of K^*S_N

where \mathbf{x} and \mathbf{m} are column vectors defined by

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad \mathbf{m} = \begin{bmatrix} m_1 \\ m_2 \\ \vdots \\ m_n \end{bmatrix} = \begin{bmatrix} E[X_1] \\ E[X_2] \\ \vdots \\ E[X_n] \end{bmatrix} \quad (6.27)$$

and κ is the covariance matrix that is defined by

$$\kappa = \begin{bmatrix} \text{Var}(X_1) & \text{Cov}(X_1, X_2) & \cdots & \text{Cov}(X_1, X_n) \\ \text{Cov}(X_2, X_1) & \text{Var}(X_2) & \cdots & \text{Cov}(X_2, X_n) \\ \vdots & \vdots & \vdots & \vdots \\ \text{Cov}(X_n, X_1) & \cdots & \cdots & \text{Var}(X_n) \end{bmatrix} \quad (6.28)$$

Eq.(6.26) shows that the pdf of jointly Gaussian random variables is completely specified by the individual means and variances and the pairwise covariance.

Now having \mathcal{M}_N and K^*S_N known, the pdf of $\hat{C}_{(S)N}$ is readily obtained by taking the linear transformation of Gaussian random variables. A very important property of jointly Gaussian random

variables is that the linear transformation of any n jointly Gaussian random variables results in n random variables that are also jointly Gaussian. If we let $X = (X_1, X_2, \dots, X_i, \dots, X_n)$ be jointly Gaussian and define $Z = (Z_1, Z_2, \dots, Z_i, \dots, Z_n)$ by

$$Z = AX, \quad (6.29)$$

where A is an $n \times n$ invertible matrix, then the pdf of Z is:

$$f_Z(z) = \frac{\exp\{-\frac{1}{2}(z-n)^T C^{-1}(z-n)\}}{(2\pi)^{\frac{n}{2}} |C|^{\frac{1}{2}}}. \quad (6.30)$$

Thus, the pdf of Z has the form of Eq.(6.26) and therefore $Z_1, Z_2, \dots, Z_i, \dots, Z_n$ are jointly Gaussian random variables with mean n and covariance matrix C :

$$\begin{aligned} n &= Am, \\ C &= AkA^T. \end{aligned} \quad (6.31)$$

We will find the pdf of $\hat{C}_{(S),N}$ by introducing auxiliary random variable. Let $Z_{aux} = KS_N$ and $Z = \hat{C}_{(S),N} = M_N + KS_N$. If we define $Z = (Z, Z_{aux})$, then

$$Z = AX \quad (6.32)$$

where

$$A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \quad (6.33)$$

and

$$X = \begin{bmatrix} M_N \\ KS_N \end{bmatrix} \quad (6.34)$$

From Eq.(6.30) we have that Z is jointly Gaussian with mean $\mathbf{n} = A\mathbf{m}$, and covariance matrix $C = A\kappa A^T$. Furthermore, it then follows that the marginal pdf of Z is a Gaussian pdf with mean given by the first component of \mathbf{n} and variance given by the 1st row - 1st column component of the covariance matrix C . If we carry out the above matrix multiplications,

$$\kappa = \begin{bmatrix} \text{Var}(\mathcal{M}_N) & \text{Cov}(\mathcal{M}_N, K\mathcal{S}_N) \\ \text{Cov}(K\mathcal{S}_N, \mathcal{M}_N) & \text{Var}(K\mathcal{S}_N) \end{bmatrix} \quad (6.35)$$

$$C = \begin{bmatrix} \text{Var}(\mathcal{M}_N) + 2\text{Cov}(\mathcal{M}_N, K\mathcal{S}_N) + \text{Var}(K\mathcal{S}_N) & \text{Cov}(\mathcal{M}_N, K\mathcal{S}_N) + \text{Var}(K\mathcal{S}_N) \\ \text{Cov}(K\mathcal{S}_N, \mathcal{M}_N) + \text{Var}(K\mathcal{S}_N) & \text{Var}(K\mathcal{S}_N) \end{bmatrix} \quad (6.36)$$

$$\mathbf{n} = A\mathbf{m} = \begin{bmatrix} E[\mathcal{M}_N] + E[K\mathcal{S}_N] \\ E[K\mathcal{S}_N] \end{bmatrix} \quad (6.37)$$

then, we find that

$$\begin{aligned} E[Z] &= E[\hat{\mathcal{C}}_{(S),N}] \\ &= E[\mathcal{M}_N] + E[K\mathcal{S}_N] \\ \text{Var}(Z) &= \text{Var}(\hat{\mathcal{C}}_{(S),N}) \\ &= \text{Var}(\mathcal{M}_N) + 2\text{Cov}(\mathcal{M}_N, K\mathcal{S}_N) + \text{Var}(K\mathcal{S}_N). \end{aligned} \quad (6.38)$$

Finally, the pdf of $\hat{\mathcal{C}}_{(S),N}$ is a Gaussian pdf with mean $m = E[\mathcal{M}_N] + E[K\mathcal{S}_N]$ and variance $\sigma^2 = \text{Var}(\mathcal{M}_N) + 2\text{Cov}(\mathcal{M}_N, K\mathcal{S}_N) + \text{Var}(K\mathcal{S}_N)$:

$$f_{\hat{\mathcal{C}}_{(S),N}}(x) = \frac{e^{-(x-m)^2/2\sigma^2}}{\sqrt{2\pi\sigma}}. \quad (6.39)$$

Fig. 6.4 depicts the pdf of $\hat{\mathcal{C}}_{(S),N}$ together with the pdf's of \mathcal{M}_N and $K\mathcal{S}_N$, when 8 Mbps, 0.1, 0.5 sec., and 2.48 are used for \bar{R} , $\bar{\rho}$, \bar{b} , and λ , respectively.

Fig. 6.5 demonstrates the accuracy of the developed analytical model, when same traffic parameters are used. The analytical model precisely captures the actual load distribution. Thus, it has been proven that that \mathcal{M}_N and $K\mathcal{S}_N$ are jointly Gaussian random variables. Another assumption, made early in this section, is also proven to be correct. We assumed that, with fairly large average number of

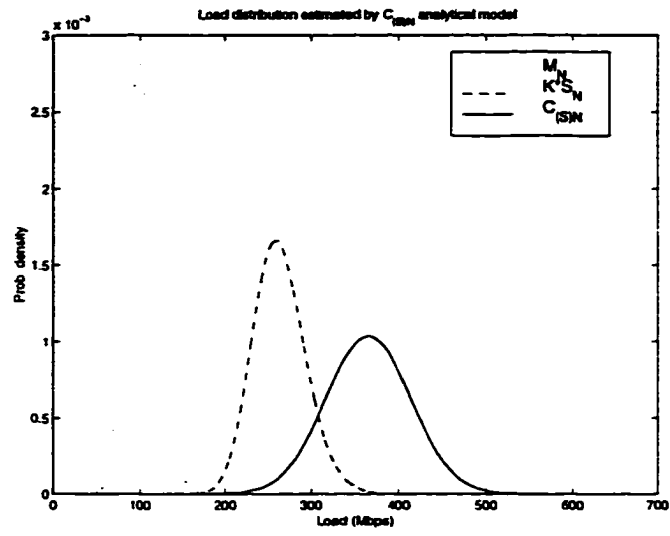


Figure 6.4 Load distribution estimated by $\hat{C}_{(S)N}$ analytical model

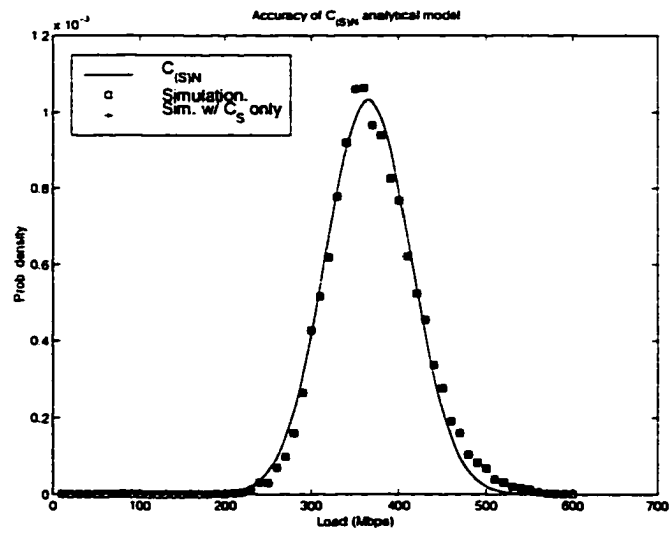


Figure 6.5 Accuracy of $\hat{C}_{(S)N}$ analytical model

connections, the equivalent capacity is largely determined by stationary approximation. The simulation results, where the stationary approximation was used alone, exactly matches with those where both approximations were used.

Load distribution when CED is used

When CED is used, input traffic is dispersed according to the estimated cost of dispersion. The dispersion cost is calculated not only based on the mean load distribution but also based on the traffic characteristics of other calls in each VP. Then the CED algorithm selects paths such that the total amount of load increased by the arriving call is minimized. As mentioned before, the load on a VP, i.e., required equivalent capacity, is not a linear function of the number of connections in the VP. Rather it is very sensitive to the traffic characteristic of individual connection. Thus, at a time instance, the load of each VP differs from that of others as shown in Fig. 6.6.

Building an accurate analytical model for CBP would require a Markov state space representation. But, even for a moderately accurate model, the state space and the complexity of state transitions explode incredibly. It makes this approach intractable. Favorably, however, long term average load distribution is nearly same as that of others in appearance. This observation brought an idea of Approximation by Single Abstract Path (ASAP).

The idea of ASAP is that an ATM switching node with N VP's where the load distribution of each VP differs, can be modeled as the one with N VP's where each VP has the same load distribution (i.e., long term average load distribution). Since, in the simplified approximation model, each VP is identical, the initial problem turns into that of finding the load pdf in a VP. We have worked out it in the previous section, where we used exponential distributions for R_p , ρ and b . However, as input traffic is dispersed to multiple VP's, the statistical characteristics of the traffic seen by each VP are much different from those of the traffic initially arrived at the multiplexer. In particular, we are interested in the distribution of R_p , ρ , and interarrival time. Without the accurate model describing the behavior of traffic as a result of CED, we would not tell the exact distributions of those traffic descriptors. Yet the approximation by the average dispersion factor (\bar{D}_f), i.e., the average number of paths taken by each call, gives us reasonable distributions close to those obtained from simulations.

Intuitively, the pdf of source utilization ρ will not be changed except when ρ is very low. This is because, with very low ρ , there might be little chance of having big dispersion gain. Fig. 6.7 confirms this postulation. In this figure, simulation data was collected at one of 8 VP's. Solid line depicts the exponential distribution with mean 0.1 which was applied initially to the arriving calls.

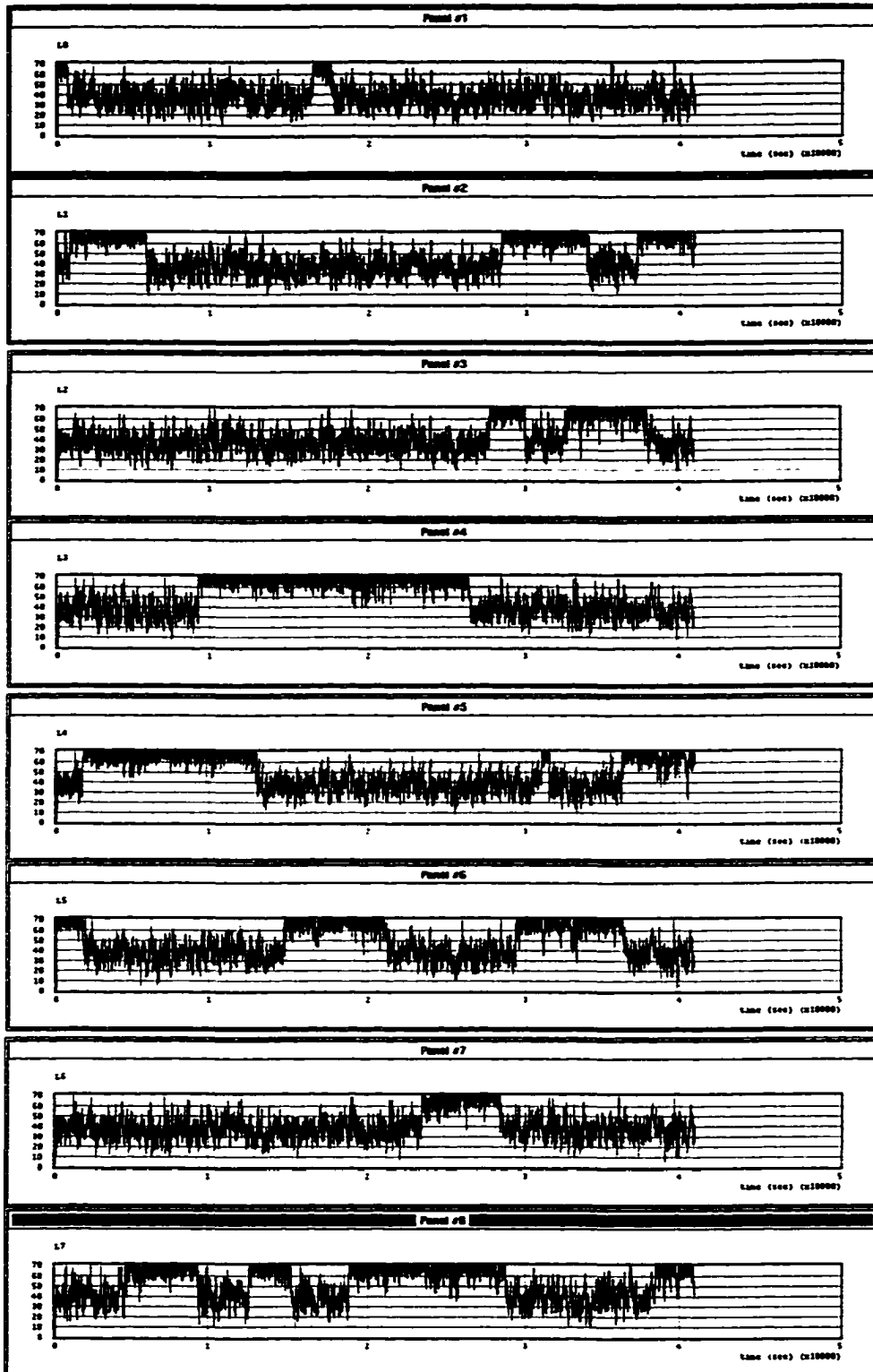


Figure 6.6 Time-varying VP loads

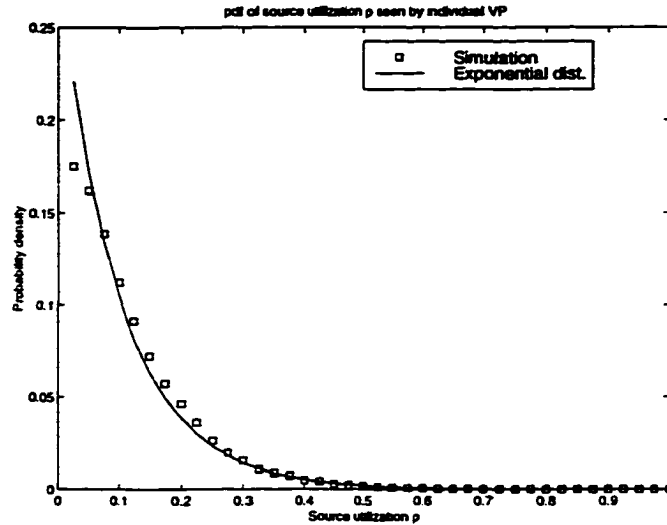


Figure 6.7 Distribution of ρ seen by individual VP

One can speculate that the pdf of interarrival time seen by individual VP will be that of exponential distribution. The only difference compared to the interarrival time of the multiplexer, i.e., the one before dispersion, will be its mean value. The mean interarrival time of connections in a VP ($\frac{1}{\lambda_i}$) can be found in a straightforward way:

$$\lambda_i = \frac{\lambda_{sys}}{N_{paths}} \bar{D}_f. \quad (6.40)$$

We assumed that each VP is selected equally likely. For example, when $R_p=8\text{Mbps}$, $\rho=0.1$, $b=0.5$ sec., and CED coef=0.05 are used in a ATM switching node with 8 VP's, for the average arrival rate λ_{sys} of 2.48, the mean interarrival time at VP_i will be 2.2247. Fig. 6.8 shows correctness of this approximation. In this calculation, 1.45 was used for the average dispersion factor \bar{D}_f , which was obtained from simulation.

The distribution of R_{peak} at an individual VP_i is altered so rigorously that it is not even close to the exponential distribution. That is because, when CED is used, R_{peak} of arriving traffic is divided evenly into multiple paths, and each path sees only one fraction of it. It is not difficult to predict that the resulting distribution of R_{peak} at an individual VP_i will be roughly a function of average dispersion factor \bar{D}_f . Yet, in detail, it is a correlated function of probability mass function (pmf) of dispersion factor and pdf of R_{peak} . The pmf of dispersion factor would not be found without solving extremely large and complicated Markov model. Instead of trying to solve such an intractable model, we would

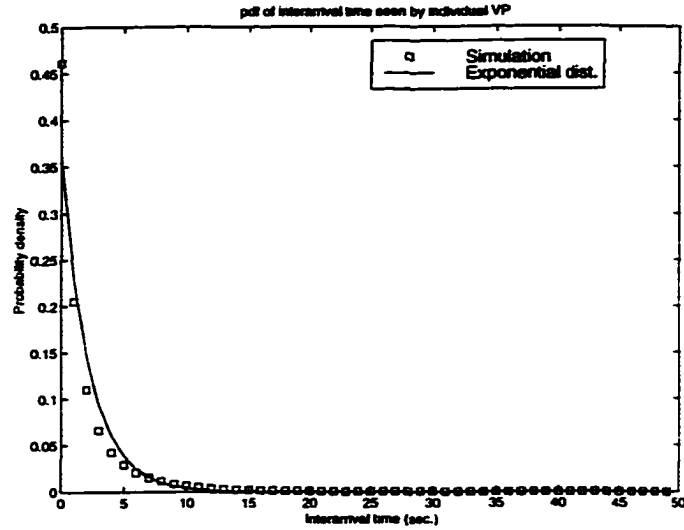


Figure 6.8 Distribution of interarrival time at a VP

like to find a good approximation using the average dispersion factor. A moderately simple estimator, which was formulated in this study, provides very accurate calculation of that quantity. We will discuss this shortly later in this section. Meanwhile, for the approximation of the peak rate distribution at the single abstract path, R_{sop} , we assume that \bar{D}_f is known *a priori*.

As illustrated in Fig. 6.9, simulation results show that R_{peak} of individual VP_i is far different from exponential distribution. Rather it looks like Gaussian distribution. This is due to the fact that traffic with high peak rate is very likely to be dispersed. As the average dispersion factor increases, the peak rates are more likely divided to more number of paths, resulting in higher probability densities at low peak rates and lower densities at high peak rates. In other words, greater the the average dispersion factor, lower the mean and the variance of the distribution. Thus, a reasonable approximation of the pdf can be a Gaussian pdf with:

$$\begin{aligned} \text{mean} &= \bar{R}_{peak} \frac{1}{\bar{D}_f}, \\ \sigma &= \bar{R}_{peak} \left(\frac{1}{\bar{D}_f} \right)^2, \end{aligned} \tag{6.41}$$

where \bar{R}_{peak} is the mean of exponential distribution which is the initial distribution applied to the multiplexer, and σ is the standard deviation for the Gaussian approximated distribution.

Compared to the pdf's of exponential distributions, these Gaussian approximated pdf's are much closer to the simulation results.

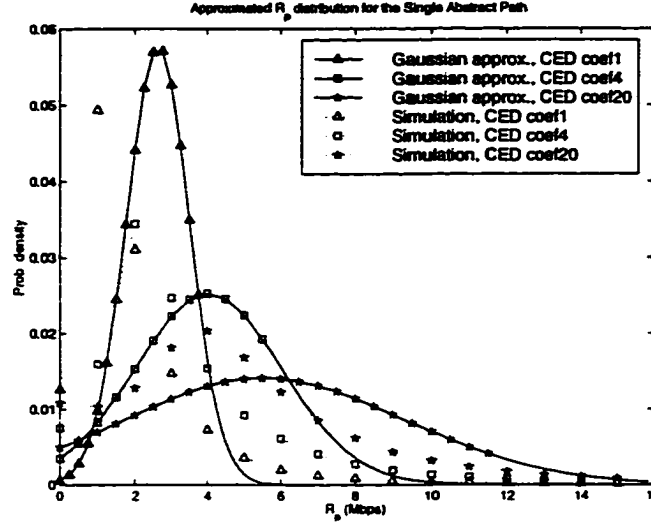


Figure 6.9 Gaussian approximation of R_{sap}

If we properly tailor the mean and the variance of Gaussian pdf, we can enhance the accuracy of approximations. Approximated Gaussian pdf's in Fig. 6.10 were obtained after multiplying scaling factors to the means and the variances calculated for the Gaussian approximations in Fig. 6.9. However, finding a general formula would not be easy, which determines precise scaling factors. Therefore, in this study, we use Gaussian approximated R_{sap} without tailoring mean and variance.

Now we know that with average dispersion factor we can find all necessary ingredients for the formulated equations in the previous section. We intend to use them to get the load distribution of the single abstract path. The only question remained is how to find the \bar{D}_f as a function of link capacity, \bar{R}_{peak} , $\bar{\rho}$, \bar{b} , mean arrival rate, and so on. By observing the simulation results, a moderately simple estimator has been formulated:

$$\bar{D}_f(x) = Ae^{(-\alpha x + \delta)} + C, \quad (6.42)$$

where x is the link capacity reserved for a VP.

Fig. 6.11 compares estimated \bar{D}_f 's with simulation results, where estimated \bar{D}_f 's are depicted by solid lines.

Changes in mean peak rates are captured exclusively by α . When mean peak rate is 8 Mbps, 0.034 is used for it in this figure. As mean peak rate increases to 16 Mbps, α decreases by half, 0.017.

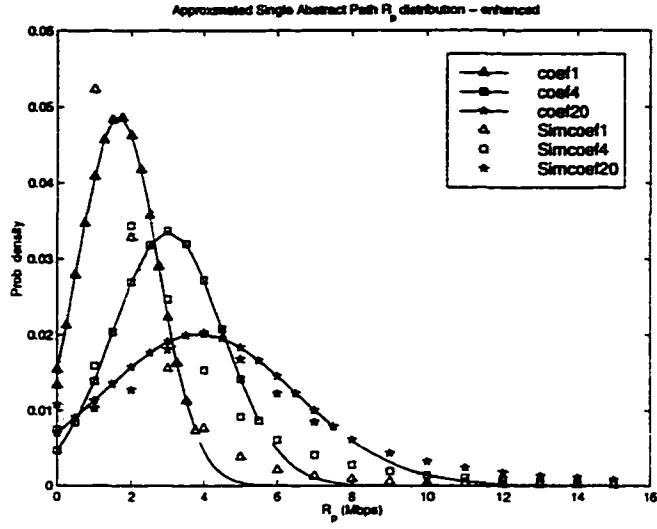


Figure 6.10 R_{sap} distribution with tailored mean and σ^2

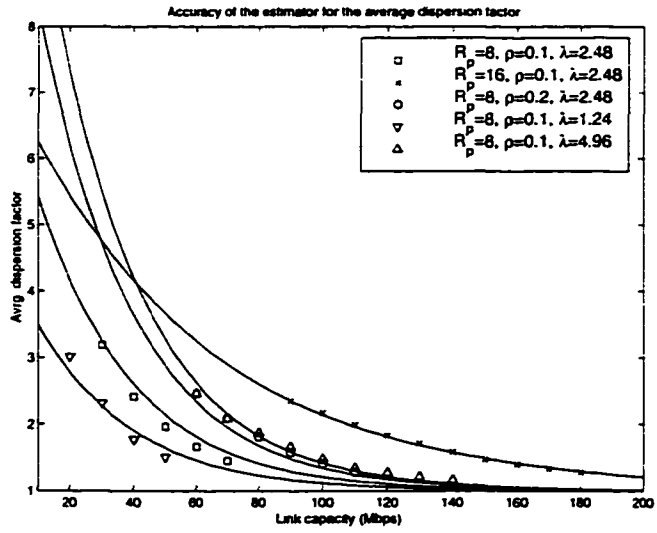


Figure 6.11 Accuracy of \bar{D}_f estimator

This makes sense because, with greater mean peak rate, the effect of increasing same amount of link capacity will be less. So it is very likely that CED will disperse more traffic. That's why the average dispersion factor decreases slowly (i.e., inducing less α) when mean peak rate is 16 Mbps. Increasing mean source utilization $\bar{\rho}$ draws increase on required equivalent capacity to some extent. However, it's effect diminishes as link capacity increases. That explains why only δ is responsible for the changes in mean source utilization $\bar{\rho}$. In this illustration, -0.25 and 0.25 were used for $\bar{\rho}$'s of 0.1 and 0.2, respectively. A is a linear function of mean arrival rate. For the λ values of 1.24, 2.48, and 4.96, A had 4, 8, 16, respectively. C is kept constant with value 1.0 for all instances in this illustration.

We have seen that, with known traffic characteristics, such as exponentially distributed R_{peak} , ρ , etc., the analytic model developed for single path provides accurate estimate of load distribution. Assuming this analytical model is accurate, we evaluate the approximations made in this section. We simplified multiple paths as identical paths using the idea of ASAP, and approximated R_{sap} as the Gaussian distribution. Other traffic characteristics, such as ρ_{sap} , λ_{sap} , and \bar{D}_f , are also approximated.

Fig. 6.12 depicts distributions of \mathcal{M}_N when 8 Mbps, 0.1, 0.5 sec, and 2.48 are used for \bar{R}_{sap} , $\bar{\rho}_{sap}$, \bar{b} , and λ_{sap} , respectively. Compared to the pdf's when CED is not used, calculated distributions are not exactly following Gaussian distributions. In the figure, each of dashed lines depicts Gaussian pdf's with mean and variance of corresponding \mathcal{M}_N distribution.

In contrast to this, distributions of $K\mathcal{S}_N$ (Fig. 6.13) are precisely the Gaussian.

Fig. 6.14 shows estimated load distributions at the single abstract path. They are not following the simulation results exactly. However, we are interested in finding CBP, and it is obtained directly from CDF of load distribution. In particular, CDF's of 90 % - 100 % are of most importance. CBP over 10 % does not have any significant meaning in practical engineering of communication networks. In that sense, this analytical model provides fairly good estimate. Note that, when CEDcoef20 is used, this analytical model overestimates the load slightly. It would not be a surprise if we consider the average number of connections. For the instances in this figure, they are 56.5, 37.1, and 27.0 for CEDcoef1 (0.0025), CEDcoef4 (0.01), and CEDcoef20 (0.05), respectively. As shown in Section 2 (in particular Fig. 2.4), for the not enough number of connections (i.e., less than 10) stationary approximation tends to overestimate the required equivalent capacity. With 27.0 average number of connections, the probability of having less than 10 connections is not negligible.

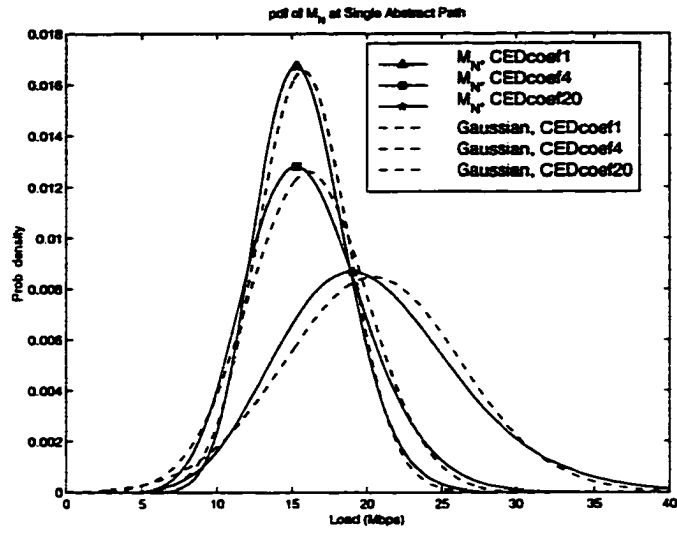


Figure 6.12 Distributions of M_N at Single Abstract Path

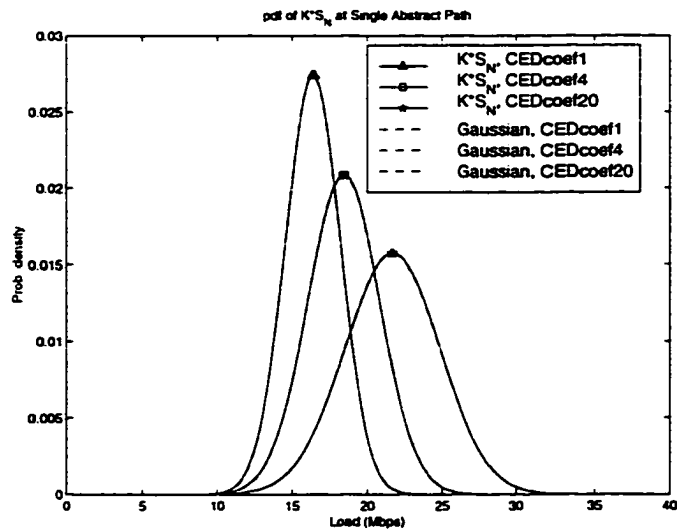


Figure 6.13 Distributions of K^*S_N at Single Abstract Path

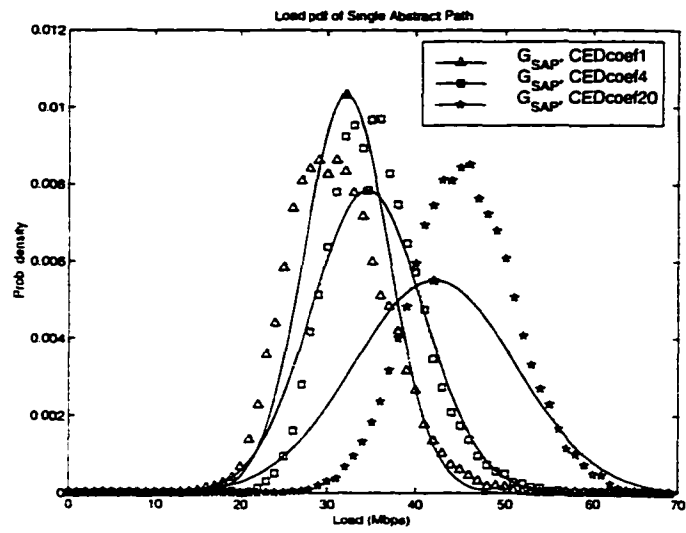


Figure 6.14 Load distributions at Single Abstract Path

7 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 a 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 algorithms including CED were 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 the CBP can be significantly reduced by CED 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. Then analytical models have been formulated. The cost effective design and engineering of VP networks requires accurate and tractable mathematical models which capture the important statistical properties of traffic. This study also revealed that the load distribution estimated by equivalent capacity follows Gaussian distribution which is the sum of two jointly Gaussian random variables. For the analysis of load distribution when cost effective traffic dispersion is used, we simplified multiple paths as identical paths using the idea of Approximation by Single Abstract Path (ASAP), and approximated the characteristics of the traffic seen by individual VP. The developed analytical models and approximations were proven to be correct when compared with simulation results.

It was shown that, in single path analysis with known input traffic characteristics, the formulated analytical model provides accurate estimate of load distribution. However, when CED is used, approximations had to be used to capture the altered traffic characteristics. As a result, although the analytical models and approximations provide fairly good estimate, it is not following the simulation result exactly. For more accurate estimate, we need to elaborate the approximations furthermore.

BIBLIOGRAPHY

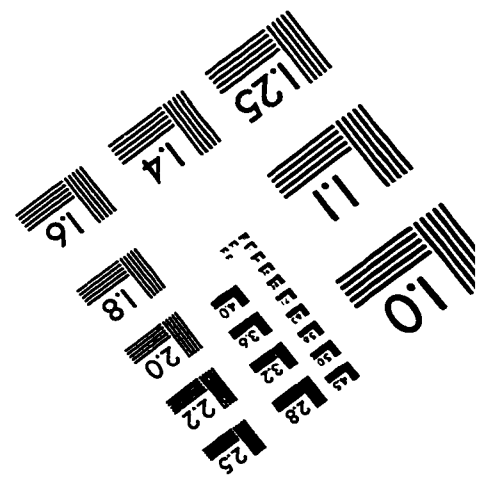
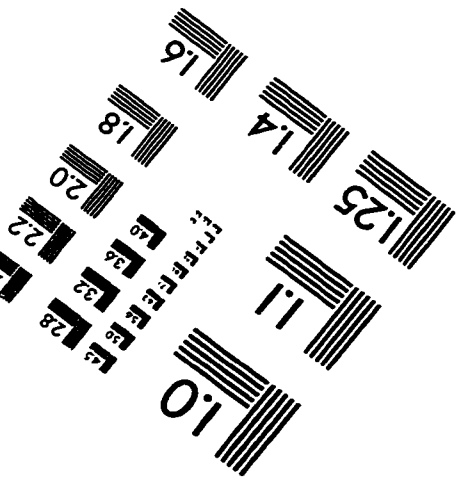
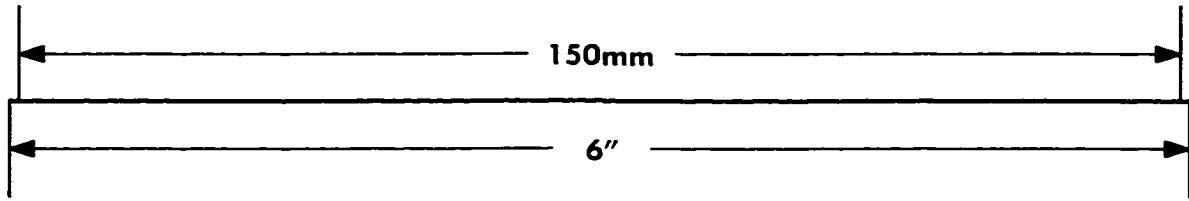
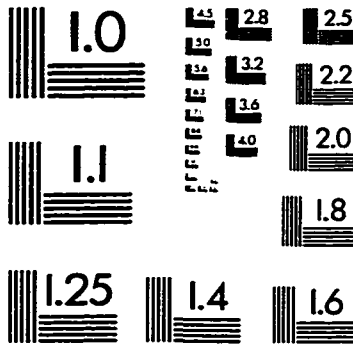
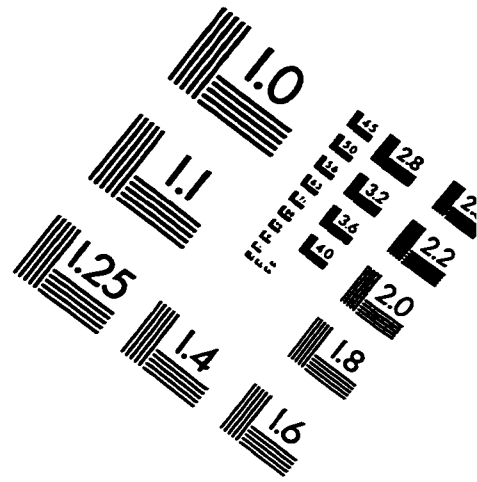
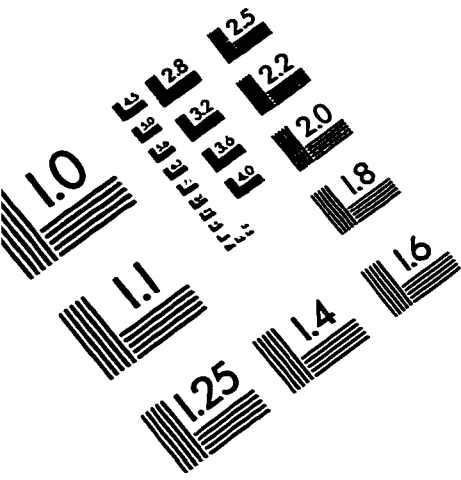
- [1] R. Guérin, H. Ahmadi, and M. Naghshineh, "Equivalent Capacity and Its Application to Bandwidth Allocation in High-Speed Networks," *IEEE Journal on Selected Areas in Communications*, vol. 9, no. 7, pp. 968–981, Sept. 1991.
- [2] E. Gustafsson and G. Karlsson, "When Is Traffic Dispersion Useful? A Study On Equivalent Capacity," in *ATM Networks: Performance Modelling and Analysis* (D. D. Kouvatsos, ed.), vol. 2, pp. 110–129, New York, New York: Chapman & Hall, 1996.
- [3] O. Gerstel, I. Cidon, and S. Zaks, "The Layout of Virtual Paths in ATM Networks," *IEEE/ACM Transactions on Networking*, vol. 4, no. 6, pp. 873–884, Dec. 1996.
- [4] Y. Sato and K. Sato, "Virtual Path and Link Capacity Design for ATM Networks," *IEEE Journal on Selected Areas in Communications*, vol. 9, no. 1, pp. 104–111, Jan. 1991.
- [5] S. Gupta, K. W. Ross, and M. E. Zarki, "Routing in Virtual Path Based ATM Networks," in *GLOBECOM '92*, pp. 571–575, 1992.
- [6] D. Anick, D. Mitra, and M. M. Sondhi, "Stochastic Theory of a Data-Handling System with Multiple Sources," *Bell System Technical Journal*, vol. 61, no. 8, pp. 1871–1894, Oct. 1982.
- [7] S. Ohta and K.-I. Sato, "Dynamic Bandwidth Control of the Virtual Path in an Asynchronous Transfer Mode Network," *IEEE Transactions on Communications*, vol. 40, no. 7, pp. 1239–1247, July 1992.
- [8] J. Y. Hui, M. B. Gursoy, N. Moayeri, and R. D. Yates, "A Layered Broadband Switching Architecture with Physical or Virtual Path Configurations," *IEEE Journal on Selected Areas in Communications*, vol. 9, no. 9, pp. 1416–1426, Dec. 1991.
- [9] R.-H. Hwang, J. F. Kurose, and D. Towseley, "MDP Routing in ATM Networks Using Virtual Path Concept," in *INFOCOM '94*, pp. 1509–1517, IEEE, 1994.

- [10] E. W. M. Wong, A. K. M. Chan, S. C. H. Chan, and K. T. Ko, "Bandwidth Allocation and Routing in Virtual Path Based ATM Networks," in *ICC '96*, pp. 647–652, IEEE, 1996.
- [11] H. G. Perros and K. M. Elsayed, "Call Admission Control Schemes: A Riview," *IEEE Communications Magazine*, vol. 34, no. 1, pp. 82–91, Nov. 1996.
- [12] 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.
- [13] A. I. Elwalid and D. Mitra, "Effective Bandwidth of General Markovian Traffic Sources and Admission Control of High Speed Networks," *IEEE/ACM Transactions on Networking*, vol. 1, no. 3, pp. 329–343, June 1993.
- [14] R. J. Gibbens and P. J. Hunt, "Effective Bandwidths for the Multi-Type UAS Channel," *Queueing Systems*, vol. 9, no. 1, pp. 17–28, Oct. 1991.
- [15] G. Kesidis, J. Walrand, and C. S. Chang, "Effective Bandwidths for Multiclass Fluids and other ATM Sources," *IEEE/ACM Transactions on Networking*, vol. 1, no. 4, pp. 424–428, Aug. 1993.
- [16] Z. Zhang and A. S. Acampora, "Equivalent Bandwidth for Heterogeneous Sources in ATM Networks," in *ICC '94*, pp. 1025–1029, IEEE, 1994.
- [17] F. P. Kelly, "Effective Bandwidths at Multi-Class Queues," *Queueing Systems*, vol. 9, no. 1, pp. 5–16, Oct. 1991.
- [18] A. Weiss, "An Introduction to Large Deviations for Communication Networks," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 6, pp. 938–952, Aug. 1995.
- [19] A. Simonian and J. Guibert, "Large Deviations Approximation for Fluid Queues Fed by a Large Number of On/Off Sources," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 6, pp. 1017–1027, Aug. 1995.
- [20] D. N. C. Tse, R. G. Gallager, and J. N. Tsitsiklis, "Statistical Multiplexing of Multiple Time-Scale Markov Streams," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 6, pp. 1028–1038, Aug. 1995.
- [21] G. de Veciana, G. Kesidis, and J. Walrand, "Resource Management in Wide-Area ATM Networks Using Effective Bandwidths," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 6, pp. 1081–1090, Aug. 1995.

- [22] C.-S. Chang and J. Thomas, "Effective Bandwidth in High-Speed Digital Networks," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 6, pp. 1091–1100, Aug. 1995.
- [23] R. J. Gibbens, F. P. Kelly, and P. B. Key, "A Decision-Theoretic Approach to Call Admission Control in ATM Networks," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 6, pp. 1101–1114, Aug. 1995.
- [24] M. Schwartz. *Broadband Integrated Networks*. Upper Saddle River, New Jersey: Prentice Hall, 1996.
- [25] V. G. Kulkarni, L. Gün, and P. F. Chimento, "Effective Bandwidth Vectors for Multiclass Traffic Multiplexed in a Partitioned Buffer," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 6, pp. 1039–1047, Aug. 1995.
- [26] W. E. Leland. "On the Self-Similar Nature of Ethernet Traffic (Extended Version)." *IEEE/ACM Transactions on Networking*, vol. 2, no. 1, pp. 1–15, Feb. 1994.
- [27] V. Paxson and S. Floyd. "Wide Area Traffic: The Failure of Poisson Modeling." *IEEE/ACM Transactions on Networking*, vol. 3, no. 3, pp. 226–244, June 1995.
- [28] E. Gustafsson and G. Karlsson. "A Literature Survey on Traffic Dispersion." *IEEE Network*, vol. 11, no. 2, pp. 28–36, March/April 1997.
- [29] E. Gustafsson and G. Karlsson. "Traffic Dispersion in ATM Networks." in *Radio Vetenskap och Kommunikation 96, RVK 96*, pp. 593–597, RVK, 1996.
- [30] E. Gustafsson and G. Karlsson. "Call Admission Control with Traffic Dispersion." in *13th Nordic Teletraffic Seminar, NTS-13*, pp. 370–383, NTNU, 1996.
- [31] E. Gustafsson and R. Ronngren. "Fluid Traffic Modelling in Simulation of a Call Admission Control Scheme for ATM Networks." in *5th Int'l Symp. Modelling, Analysis and Simulation of Comp. and Telecom. Sys., MASCOTS '97*, pp. 110–115, IEEE, 1997.
- [32] J. H. Déjean, L. Dittmann, and C. N. Lorenzen, "String Mode-A New Concept for Performance Improvement of ATM Networks," *IEEE Journal on Selected Areas in Communications*, vol. 9, no. 9, pp. 1452–1460, Dec. 1991.
- [33] S. Chowdhury, "An Analysis of Virtual Circuits with Parallel Links," *IEEE Transactions on Communications*, vol. 39, no. 8, pp. 1184–1188, Aug. 1991.

- [34] A. Jean-Marie and L. Gün, "Parallel Queues with Resequencing," *Journal of the Association for Computing Machinery*, vol. 40, pp. 1188–1208, 1993.
- [35] A. Leon-Garcia, *Probability and Random Processes for Electrical Engineering*. Reading, Massachusetts: Addison-Wesley, second ed., 1994.

IMAGE EVALUATION TEST TARGET (QA-3)



APPLIED IMAGE, Inc
 1653 East Main Street
 Rochester, NY 14609 USA
 Phone: 716/482-0300
 Fax: 716/288-5989

© 1993, Applied Image, Inc., All Rights Reserved