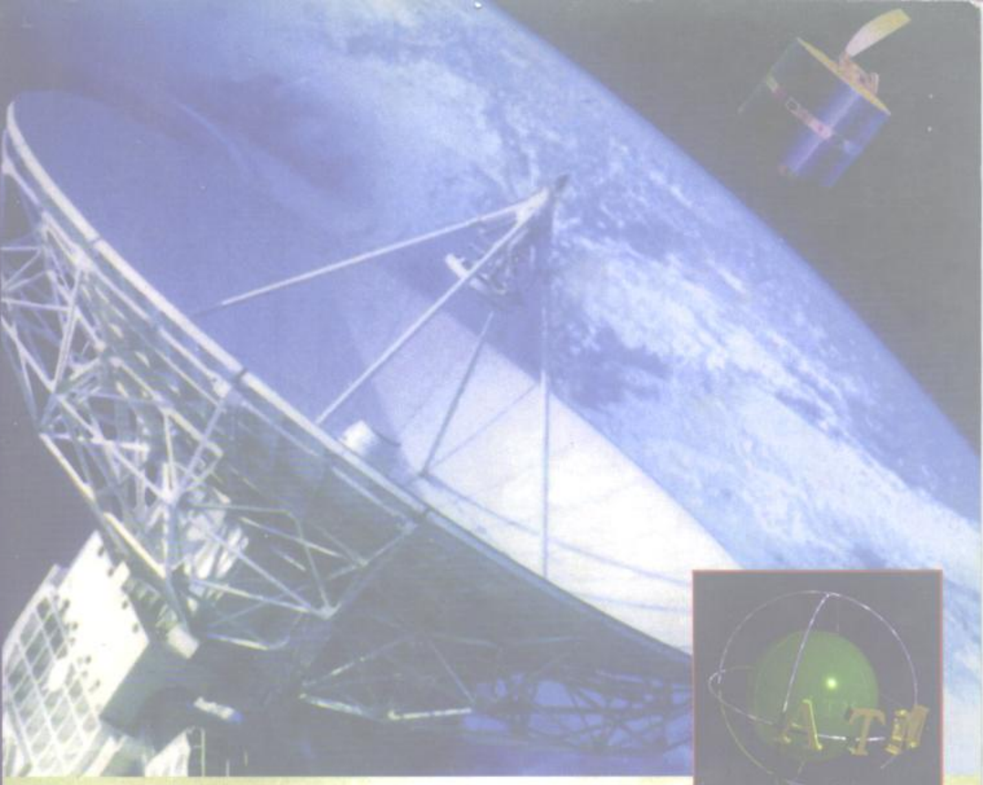
ATM与B-ISDN技术丛书

版）



［美］MISCHA SCHWARTZ

BROADBAND INTEGRATED NETWORKS

宽带网络性能分析

清华大学出版社
http://www.tup.tsinghua.edu.cn

PRENTICE HALL

BROADBAND INTEGRATED NETWORKS

宽带网络性能分析

Mischa Schwartz

清华大学出版社
Prentice-Hall International, Inc.

(京)新登字158号

图书在版编目(CIP)数据

宽带网络性能分析:英文/(美)施瓦茨(Schwart z,M.)编著.—影印版.
—北京:清华大学出版社,1998.1
(ATM 与 B—ISDN 技术丛书)
ISBN 7-302-02818-4

Ⅰ.宽… Ⅱ.施… Ⅲ.①宽带通信系统-异步通信网-网络分析-英文②宽带通信系统-分析-英文 Ⅳ.TN914.4

中国版本图书馆 CIP 数据核字(98)第 01437 号

出版者:清华大学出版社(北京清华大学校内,邮编 100084)
　　　　http:// www.tup.tsinghua.edu.cn
印刷者:北京市清华园胶印厂
发行者:新华书店总店北京科技发行所
开　本:850×1168 1/32　印张:12
版　次:1998 年 4 月第 1 版　1998 年 7 月第 2 次印刷
书　号:ISBN 7-302-02818-4/TN·87
印　数:4001～9000
定　价:19.80 元

出版前言

90 年代中期掀起了信息高速公路的浪潮。宽带综合业务数字网络(B-ISDN)代表着国家信息基础设施的最高网络层次，将在下一世纪发挥非常重要的作用。ATM 是 B-ISDN 的核心技术，已经得到了迅速地发展。广大科技人员和大专院校的师生为了掌握该领域最新发展的知识，迫切需要一套全面、系统地介绍 ATM 与 B-ISDN 详细技术的文献，为此我们精选了一些最新英文版图书，组成一套《ATM 与 B-ISDN 技术丛书》，影印奉献给广大读者。

本套丛书既系统全面，又分工明确，各有侧重。在内容安排上包括 ATM 与 B-ISDN 技术基础、宽带网信令、宽带网性能分析、ATM 网的规划与管理、ATM 网与其它网的互通以及 ATM 网络的应用等技术。希望这套丛书对从事 ATM 和 B-ISDN 研究的广大科技人员和大专院校师生有所帮助。

清华大学出版社
Prentice Hall 公司

1998.4

Preface

Activity in the area of broadband integrated networking has been expanding at a rapid rate. Commercial ATM switches, particularly for the LAN market, have gained wide-spread acceptance and have been deployed in many laboratories and other establishments. The ATM Forum has been attempting to expedite the development of standards for ATM-based networks so that vendors can produce products that will interwork compatibly with one another. The volume of papers published, covering everything from traffic characterization for this new field to performance studies of proposed control mechanisms to proposals for improved ATM switch designs, appears to be increasing as well. This makes it difficult for newcomers to the field to quickly become familiar with, or master, aspects of the field of interest to them.

Survey and tutorial papers, covering various aspects of broadband ISDN and ATM technology in a descriptive and qualitative manner, have appeared in technical journals and magazines from time to time. Books on the subject, written in a similar descriptive manner, have begun to appear as well. These books and survey papers can be used to address the problem noted above. There is still a need, however, for a textbook that provides an introductory, quantitative approach to the study of broadband integrated networks. This book is designed to fill this need. Its focus is on modeling and performance analysis in the high-speed ATM networking environment. As noted above, the literature covering quantitative design and performance issues in ATM is vast and is growing rapidly. A reader mastering the material presented here should have no difficulty understanding literature in the field. It is the author's firm belief that with the quantitative understanding comes a much better appreciation of the design issues involved in developing new products or deploying them, when available. This has always been the case for new technology, and the ATM networking world is no exception.

Because the book stresses modeling and performance issues, it requires some knowledge of elementary queueing theory. (More advanced concepts are introduced within the text where needed.) For those readers not familiar with the basic aspects of queueing theory, or those requiring a quick review, an appendix is provided that introduces the reader to the material needed to begin studying the quantitative portions of the book.

Most of the book has been tested in class. It is based on, and is an outgrowth of, a set of notes developed for a graduate course on broadband integrated networks taught at Columbia University for a number of years. The notes were also used by the author to present a series of weekly lectures on the subject during a sabbatical with the Electronic and Electrical Engineering Department of University College London.

The course given at Columbia had, as a prerequisite, an introductory quantitative course on computer networks, which introduced the student to the necessary queueing theory and analysis noted above. In universities where no such prior course is given, or where a prior course in queueing theory is not required, a course covering the material in this book could be offered with the material in the appendix presented first. The book thus lends itself to a variety of courses. Because of the quantitative nature of much of the material, however, a working knowledge of probability theory is a definite prerequisite.

The book begins with an overview of the types of services expected to be provided over ATM networks, and the resultant traffic types that might be expected to use these networks. These include voice, video, images, files, and bursty data, among other types. It then discusses the ATM protocol model, with emphasis given to the ATM layer and the various ATM Adaptation Layer (AAL) types designed to fit above it. This descriptive introduction to broadband ISDN and ATM in Chapter 2 is followed by the first quantitative chapter, Chapter 3, which provides an introduction to traffic characterization. Two types of models used to represent traffic are stressed in this chapter: fluid source modeling and the Markov-modulated Poisson Process (MMPP). Packet-voice modeling, using these two techniques, is introduced first because of its relative simplicity and because packet voice is relatively well understood. The same techniques are then applied to modeling variable bit rate (VBR) video sources. These two types of models are then combined in presenting the well-known cell and burst regions occurring in determining buffer loss probability as a function of traffic load. The chapter concludes with a brief discussion of the two-state Markov process, a special case of the Markov-modulated process, as a useful way of representing image traffic or intermittent file transfers.

Note that the stress here is not on providing an encyclopedic overview of various models proposed for different traffic types. The field of traffic characterization is, by itself, a very active and vital one, with the number of papers on the subject proliferating rapidly. The treatment here is introductory and tutorial only. The author has, of course, demonstrated a certain personal bias in selecting models with which he is familiar, or which he has personally found interesting and useful. Workers in the field may feel their favorite models have been slighted. The only answer the author can give is that a reader following the discussion here should have no difficulty reading the current literature and deciding for himself/herself which particular model should be most appropriate for some particular purpose. Much of the rest of the book follows the same pattern. The author has selected studies or topics of most interest to him. Collectively, however, they do provide a thorough introduction to the quantitative design and performance issues arising in broadband ATM networks carrying integrated traffic. A reader should thus have no problem continuing with self-study in the area.

Chapter 4 uses the two models mentioned (with particular stress on the two-state Markov process for simplicity) to study admission and access control in broadband networks. Access control here is limited to the leaky bucket technique which has become almost a classic control procedure. A reader absorbing this material should again have no problem in studying and evaluating other control techniques proposed in the literature. Chapter 5, on ATM switches, stresses the significance of using output queueing where possible, and develops the now-classic penalty introduced by using input queueing. It provides examples of shared memory, shared medium, multistage space, and self-routing switches. Comparisons are given where possible.

Chapters 3 through 5 focus on performance analysis at one node in an ATM network. Chapters 6 and 7 attempt to expand the performance study to a path in the network. This is an extremely difficult area of study, with much research currently underway. For this reason, Chapter 6 covers bounds of performance only. As a byproduct, the concept of effective capacity, currently a very "hot" topic, falls out very nicely. There are tradeoffs in the use of this concept: The impact of sources on admission control can be evaluated simply and additively, but the resultant control policies are quite conservative, with the advantage gained by statistically multiplexing sources in a network often absent. Finally, Chapter 7 discusses and analyzes feedback control of congestion in a high-speed network. One section is devoted to the ATM Forum proposal for a rate-based control mechanism; another section covers an interesting proposal for window control in a large delay-bandwidth environment, characteristic of the wide-area, high-speed ATM networks coming in the not too distant future.

The field covered here is vast. As already noted, much selectivity had to be exerted to keep this book at a reasonable size. Important topics had to be left out out of necessity. Rapid improvements in the technology and better understanding of the characteristics of the traffic transmitted over broadband integrated networks once deployed may change some of the conclusions described in the book. But the main point, as already noted, is not to provide an encyclopedic description of models and techniques, possibly useful in the design of these networks, but to provide a clear, tutorial introduction to modeling and performance analysis, so that a reader will be able to understand and keep up with new developments as they arise. The author can only hope he has been successful in this goal.

Mischa Schwartz
New York

Acknowledgments

As is always the case with writing textbooks of a more advanced nature, such as this one, there are many individuals whose help the author must acknowledge. First and foremost are the students who took the graduate course for which the book was developed. Their questions and responses to queries, as well as projects carried out as part of the course requirements, forced the author to think issues through clearly and present ideas and concepts in an understandable way. He has found, over the years, that presenting this type of material to a discerning audience is the best way of gaining a clearer understanding of the subject for himself. Two of his doctoral students, Paul Skelly, now at GTE Laboratories, and Ness Shroff, now at Purdue University, carried out original research reflected in a number of the sections in this book. Regular interactions and discussions with them also forced the author to come to a better understanding of much of the material presented here. He thanks both of them for the use of material appearing in their doctoral theses.

Keeping up with the rapidly changing standards and standards proposal discussions at the ATM Forum and ITU-T, as well as keeping abreast of products reaching the marketplace, is very time-consuming and difficult for an academic to carry out. One must thus rely heavily on input from individuals in industry for this information. The following individuals were particularly helpful in answering questions and providing material to use in various places in the book: Wai Chen and Faramak Vakil of Bellcore; Kai Eng, Xiaoqiang Chen, and Carolyn Ngueyen of AT&T Bell Laboratories; and Giovanni Pacifici, of the Center for Telecommunications Research (CTR) at Columbia.

A pioneering project on a light-wave-based, very high-speed network of the future called Acorn (carried out at CTR some years ago jointly between CTR students, staff, and faculty, and many industrial representatives) stimu-

lated the author into thinking much more deeply about the area of high-speed integrated networks. This project was conceived by Professor Anthony Acampora, the then-Director of CTR. The author had the fortunate opportunity to head up the Network Management and Control Group for this project, and thanks members of that group for the many discussions of control issues arising at meetings of the group. Subgroups led by industrial representatives were organized to suggest ways of implementing access and admission control on the Acorn testbed.

The author would also like to acknowledge support for his research on broadband integrated networks from the U.S. Office of Naval Research (ONR) and from the National Science Foundation through its support of CTR, one of the original Engineering Research Centers. This support was absolutely essential in setting the groundwork for developing the graduate course for which this book was written. Finally, the author again thanks Mrs. Betty Lim, Editorial Assistant of the Electrical Engineering Department at Columbia, for preparing and producing the text and accompanying figures for, first, the notes, and then the book manuscript. Her unfailing help has been invaluable.

Contents

Chapter 1

Introduction

Most existing telecommunication networks encompass either one of two types: the ubiquitous circuit-switched networks (characteristic of telephony) carrying principally voice traffic; packet-switched networks used primarily to transmit data of various types [SCHW 1987]. Telephone networks have been with us for over a hundred years. The packet-switched data or computer communication networks are much more recent, having first been deployed in the late 1960s.

The digitization of most of the public telephone networks worldwide, as well as the concurrent deployment of optical fiber, now allow potentially much wider bandwidths (higher bit rates) to be used than has previously been the case for the voice and data networks mentioned above. This has, in fact, been universally recognized by the adoption of worldwide standards for very high bandwidth digital transmission over optical fibers. The standards define a hierarchy of synchronous time-multiplied transmission, compatible with optical transmission [ANSI 1991a] [SCHW 1990]. In North America, the digital hierarchy, called SONET, ranges from 51.84 Mbps to 2.48832 Gbps (and poten-

[SCHW 1987] Schwartz, M., *Telecommunication Networks: Protocols, Modeling and Analysis*, Reading, MA: Addison-Wesley, 1987.

[ANSI 1991a] *American National Standard for Telecommunications Digital Hierarchy Optical Interface Rates and Formats Specifications*, ANSI T1.105-1991, New York: American National Standards Inst., 1991.

[SCHW 1990] Schwartz, M., *Information Transmission, Modulation, and Noise*, 4th ed. New York: McGraw-Hill, 1990.

tially even higher) in multiples of 51.84 Mbps. The equivalent CCITT[1] international recommendation for a Synchronous Digital Hierarchy (SDH) joins SONET at the higher transmission rates, but differs somewhat at the low end of the hierarchy because of the need for compatibility with existing digital telephone transmission systems at 50 Mbps and below.

With such bit rates as 155 Mbps, 622 Mbps, and 2.4 Gbps becoming available, it is natural to start planning new user services that would make use of these bandwidths. Prime candidates include applications involving high-resolution images and video. In addition, it becomes natural to start thinking of merging services currently provided by different networks (e.g., voice and packet data) onto one common network. This desire to *integrate* services is being pushed, as well, by the computer venders' move toward multimedia workstations and the transport capability needed to support multimedia communications.

Activity has thus begun in earnest, worldwide, to develop *B*roadband *I*ntegrated *S*ervices *D*igital *N*etworks (B-ISDN) capable of carrying out the functions just described. This activity is taking different forms. The International Consultative Committee on Telephony and Telegraphy (CCITT), now renamed the International Telecommunications Union-Telecommunication Standardization Sector (ITU-T), has issued a number of Recommendations (its terminology for standards documents) concerning B-ISDN. Since much of the traffic projected to be deployed over these networks is new and not yet supported over any digital network, the potential traffic must be characterized in order to properly design such integrated networks. Much research in the past few years has in fact been devoted to studying and characterizing the potential video and image traffic, as well as studying ways of controlling its impact on networks. Finally, manufacturers have been developing and have begun to provide switches capable of being deployed in broadband integrated networks.

Actively involved in standards development worldwide is the ATM Forum, a voluntary group of several hundred manufacturers, vendors, communication carriers, and other organizations with interests in seeing ATM standards development speeded up to expedite the delivery of ATM products to the marketplace. Since the ATM Forum has no official status with the ITU-T the standards recommendations made by this body can be unofficial only. However, because of the worldwide influence of the companies participating in its activities, its recommendations are equivalent to defacto standards. We shall have occasion to refer to some of the ATM Forum recommendations in this book.

[1] CCITT was renamed ITU-T, the International Telecommunications Union-Telecommunication Standardization Sector, in 1993. Both terms will be used in the material following, since reference will be made to older CCITT documents as well as to newer ITU-T Recommendations.

In this book we first summarize the current state of the CCITT Recommendations for B-ISDN. We focus principally on *ATM* (Asynchronous Transfer Mode), a cell-based (short, fixed-length packet) mode of transmitting integrated traffic through broadband networks. We then move to the topic of traffic characterization in Chapter 3 and study, comparatively, a number of models proposed for characterizing traffic. We begin with voice, since this traffic type is well understood, although its representation in packet form is relatively recent. Some of the models proposed for packet voice are then carried over to video characterization, an area still under intense study. We conclude the topic of traffic characterization by briefly introducing one model often used to represent image or data file traffic in an interactive mode.

The discussion in Chapter 3 is quantitative and relies on some prior knowledge of elementary queueing theory. A brief introduction to this subject appears in the Appendix for those readers not familiar with the subject or for those desiring a review. More advanced topics are introduced both in this chapter and later in the book when needed. The analytical techniques and traffic models developed in this chapter are also used in later chapters when needed.

Given a number of ways of characterizing the traffic that might be deployed over broadband networks, it is then natural to use these models to study traffic control into, and across, the network. Control is required because of limited capacity within the network. It is thus important to ensure that no one user dominate the network, that each user traffic source be provided the appropriate quality of service required, and that the aggregate traffic not congest the network. The Quality of Service (QoS) defined for each traffic class is, in these networks, a multidimensional performance objective function including such parameters as delay (mean and/or specified quantile), call blocking probability, cell (packet) dropping probability, end-to-end time jitter, throughput, and the like.

Traffic control takes at least three forms. First, there is admission control—a decision must be made whether a user, desirous of establishing a connection (call) over the network, can be accommodated. This is based on some estimate of the characteristics of the traffic to be transmitted. Second, once a call is admitted, control must be maintained at the entrance (access) to the network to ensure the traffic entering the network receives its negotiated quality of service and does not adversely affect other traffic. This is done by appropriately scheduling the different traffic classes and elements within them and by maintaining a so-called "policing" function to ensure each user abides by its traffic estimate. Admission and access control are discussed in Chapter 4.

Third, control must be maintained throughout the network and at the various traffic destinations to ensure network and receiver congestion does not develop, or, if it starts developing, to quench it as quickly as possible. Network congestion control mechanisms are not new. They have been studied and used for years in both circuit-switched and packet-switched networks [SCHW 1987]. What makes the admission, access, and congestion control problem particularly different in the broadband integrated environment is the variety of traffic types (classes) to be accommodated (each to be provided with a potentially different quality of service) and the very high transmission bandwidths (bit rates) of these networks. Because of the high bit rates, the transmission times of cells (packets) being switched through the network are very short. Thousands of them may thus potentially be enroute between any two source-destination points in even a moderately sized, wide-area network. The problem of preventing or responding to congestion in a network thus takes on dimensions never before encountered in telecommunication networks. This is a point we will be particularly stressing in our discussion of congestion control later in Chapter 7.

Following the discussion of network traffic control in Chapter 4 we move into the realm of switching for B-ISDN in Chapter 5. A number of high-speed ATM switches have been proposed; some are available in the market. They can be categorized in a number of ways. We carry out a comparative study of simple models of some of these in Chapter 5, focusing on their complexity and throughput/delay characteristics.

The traffic characterization and admission/access control studies in Chapters 3 and 4, respectively, focus on representation at one point in a network, normally at the entrance point, or user-network interface. The characteristics of a given traffic stream change, however, as the stream progresses through a network, due to queueing and interaction with other traffic streams at switching and buffering points. This makes the study of traffic end-to-end across a network quite complex. In addition, admission control, as we shall see, should be carried out considering the impact of a newly admitted call of any type on previously established connections along its entire proposed path end-to-end. The end-to-end properties of traffic are thus vital to a realistic analysis of its impact on a network.

In Chapter 6 we introduce this subject by summarizing work on end-to-end performance bounds that has appeared in the literature. This study leads quite naturally, as we shall see, to a concept called the *equivalent* or *effective capacity* of a given traffic source that has received a lot of attention. This property of a source is additive, enabling us to easily estimate the effect of a given source on other sources already present at a given multiplexing point in a network. This in turn simplifies the admission control problem at a given entrance point. The use of this property does result in an overly conservative control in some cases since the statistical multiplexing or smoothing advantage known to occur when sources are combined at a point is ignored.

Chapter 7 returns to the problem of network control, focusing on feedback control mechanisms devised to control network and receiver congestion. Both window and rate control techniques are discussed, with prime consideration given to the impact of high-speed transmission on the control mechanisms. The distinguishing characteristic here, as contrasted to lower-speed packet switching, is that the propagation delay from source to destination plays a key role. Questions of stability arise in wide-area broadband networks, particularly where the propagation delay is many times the packet or cell length. Alternately put, the control in this case has to be designed with the knowledge that many packets may be enroute over a given path end-to-end, and cannot simply be dropped when congestion is encountered.