A Probability Theory 457

A.1 Some Functions Used in Probability Models 457
Binomial Coefficients and Related Polynomials 457
The Kosten and Brockmeyer Polynomials 459
Moments of a Distribution 459
Generating Functions and Transforms 460

A.2 Birth-and-Death Processes 461

A.3 Renewal Processes and the $GI/M/N/\infty/N$ Queue 462
Distribution of Arrivals in a Fixed Interval 462
Distribution of Interarrival Times in a Random Interval 466
Compound Process in the $GI/M/N/\infty/N$ Queue 467
References 468

B Optimization Theory 469

B.1 Convexity 469
Convex Sets 469
Convex Functions 470

B.2 Optimality Conditions for Optimization Problems 471
First-Order Conditions 472
Second-Order Conditions 473
Global Optimality 473

B.3 Duality Theory 474

B.4 Numerical Methods for Unconstrained Problems 475
Line Search 476
Direction Calculation 477

B.5 Numerical Methods for Constrained Problems 480
Primal Methods 480
Relaxation Methods 484

B.6 Linear Programming 487
Linear Programming Duality 490
Network Flows 491
References 493

C Notation and List of Symbols 494

C.1 Notation 494

C.2 List of Symbols 495

D Numerical Data 498

E Answers to Some Problems 506
References 538

CHAPTER 1

Introduction

Since the early 1960s, the world of telecommunications has undergone dramatic changes, which will continue to accelerate. Twenty-five years ago, there were only two distinct telecommunication networks, each geared to a particular type of service. The telegraph network transmitted information in digital form, operated in a store-and-forward manner via message-switching techniques, and was dedicated to the transmission of written information. The telephone network transmitted information in analog form, operated in a circuit-switched mode, and was used exclusively to transmit voice. Each network served a well-identified need, and there was relatively little interconnection between them. As with many other technology-dependent areas, telecommunications started to change with the arrival of computers in the early 1960s.

Computers had a double impact on telecommunications. First, the communication requirements between computers themselves created a new class of user that was not well served by either of the existing networks. The transmission of digital information requires a low probability of error, which the analog telephone network was not designed to provide. Also, interactive computer use entails relatively long periods of inactivity between requests for transmission, which means that the line is not used efficiently and can remain idle for a substantial fraction of a call's duration. The digital telegraph network, on the other hand, was too slow for computer communications.

At the same time, it became obvious that in addition to using telecommunication networks, computers could be used within these networks to carry out many of the management and planning functions required to operate them smoothly — in particular, switching messages within the network. With the combination of computers and the increased use of digital transmission, there soon emerged a new class of telecommunication network designed specifically for the need of a new type of customer, the remote interactive computer user. The growth of these specialized networks has been dramatic; within the last few decades, they have appeared in all industrialized countries.

Telephone networks have been similarly transformed. Most of the switching is now done by computers, as opposed to wired-logic electromechanical devices, and transmission systems are increasingly being converted to the digital format.

While the dichotomy between the voice and data networks still exists, the distinction is becoming less sharp. Both networks use digital transmission methods, at least in the transport network; in both, computers are responsible for switching information through the network. This trend is becoming more pronounced as the demand for digital services such as facsimile increases and as new, high-capacity services such as low- or high-speed video services are introduced. Since all these services will be digital, the concept of a unique network carrying all types of information in the form of bit streams immediately comes to mind; clearly it is only a matter of time before the two networks are merged into one.

In the 1990s network planners will be engaged in designing such integrated networks. Because planning techniques are evolutionary in nature, the new planning techniques will be based on current methods, modified to handle the new problems arising in an integrated environment. Thus it is necessary to understand the methods developed for the existing networks before embarking on the development of new tools for more complex systems.

Planning telecommunication networks is a complex process, drawing on the techniques of electrical engineering, economics, probability theory, and operation research. Many elements of this process, such as forecasting or transmission engineering, are not specific to telecommunication networks, but occur in many other areas of engineering or economics. One specialized part of planning, however, is closely related to the particular nature of the network: *traffic engineering* as it is traditionally called. Because it is so specialized, traffic engineering is not treated outside the telecommunications area.

The telephone network has been in existence for nearly a century, and there exists a large body of information on the subject of modeling and planning these networks. This information, however, largely confined to telephone companies, is reported at little-publicized conferences such as the International Teletraffic Congress. In comparison, a very large body of information on models and planning techniques is available about data networks, although they have existed for less than 25 years. This information is widely published in the open technical literature. This book collects and synthesizes the known planning methods that are specific to the traffic engineering of circuit-switched networks.

The emphasis is on mathematical models, not only with respect to the current network, but, more importantly, with a view on potential generalizations applicable to the integrated network case. First, however, we must describe the contents of the book in relation to other areas of telecommunication network planning. We do this from two points of view. In Section 1.1, we describe a standard hierarchical model of telecommunication systems, using it to show where the problems addressed in the book appear in this hierarchy. In Section 1.2, we relate this model to current switching techniques, and in Section 1.3, we consider the planning of networks in general, showing at which

points in this process it is important to take into account that we are dealing with a telecommunication network operating in a particular mode.

1.1 The OSI Layered Architecture

The transmission of information between widely separated points is a complex process that requires the cooperative operation of a large number of systems. Designing and implementing these systems is much too difficult to be tackled in a single stage, and a systematic approach is required. The organization of cooperating processes is not specific to the world of telecommunications, but is found in computer science, operations management, and engineering, among other areas. There are many ways to organize cooperating processes efficiently to achieve some end; although the centralized approach is conceptually the simplest, other organizations such as sequential or network modes are also used.

One type of organization is based on the notion of a layer. Briefly stated, a layer at level n is defined by a set of processes called *peer processes*, a set of functions that can be performed by these processes, and by a set of rules, also called *protocols*, that govern the actions in question. Each layer is responsible for a precise subset of the operations required to perform the global task. These functions are implemented by sending requests for service to lower layers, which provide services to the layers above them. The usefulness of the layered approach comes from the requirement that the operation of a layer must be independent of the other layers; the implementation must also be independent from one layer to the next.

In general, the actions and the protocol are defined at level n in such a way that the user has no indication as to how these functions are actually carried out. This specification is done in terms of functions of layer $n - 1$, but is of no interest to the user of layer n, and in fact is carefully hidden from him or her. Hiding the implementation of the function — an essential feature of the layered approach — offers a measure of independence between processes that would otherwise be seen as strongly coupled.

As an example, consider the user of electronic mail on a computer. The processes could be the mail processes, with the corresponding actions of reading, sending, and copying mail. The protocol for this layer is a set of rules that specifies how the messages are to be organized, for instance, the format of the addresses and the meaning of the sender, recipient, and copy fields. The only concern of the mail user is to construct each mail message according to these rules. Once the message is ready, the user requests that it be sent to its destination. The layered architecture guarantees that whatever operations are required to send the mail are totally invisible to the mail user. The user

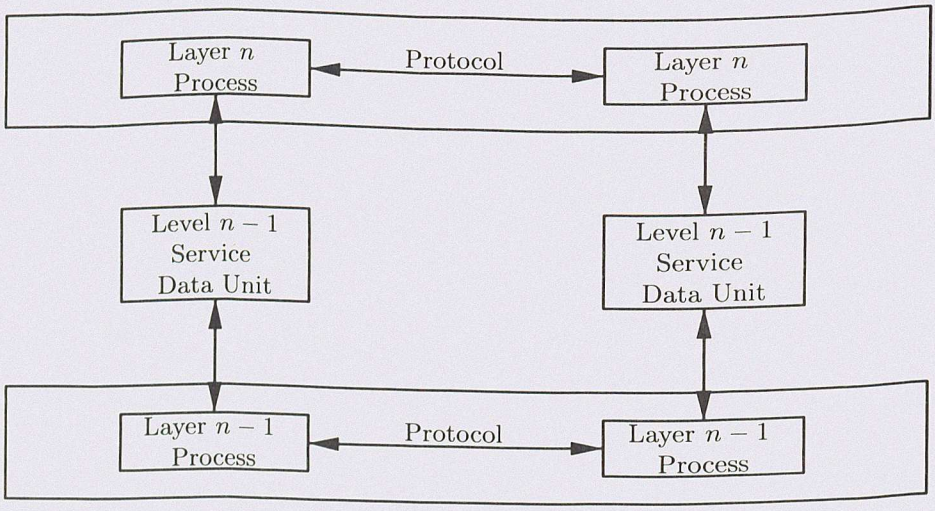

Figure 1.1 The Layer Organization of Cooperating Processes

has no knowledge about the way the message is transmitted to the destination, whether by a virtual connection over a public packet-switched network, a telephone connection over the public telephone network, or some other means. The obvious advantage of masking functions in this way is that the mail-level functions remain unchanged even if the transport function is implemented differently.

To repeat, there are two types of communication in the layered approach. The communication between peer processes is regulated by the protocol at the level at which the process operates. The communication is logical in the sense that the exchange of information between the peer processes, although appearing to be a direct exchange, is actually carried out by a series of requests to lower levels. This communication between adjacent levels is done via rigidly defined formats called *service data units*, which serve to mask implementation issues to the user of level n. This distinction is illustrated in Fig. 1.1 for levels n and $n-1$.

The *International Standards Organization* (ISO), in collaboration with the *Comité Consultatif pour le Téléphone et le Télégraphe* (CCITT), has sponsored studies in defining such a layered model applicable to telecommunication systems. This model is known as the *Open System Interconnection* (OSI) model. Although not the only one in use, the OSI model provides a good basis for discussing various communication techniques. It includes most of the features found in commonly used systems such as DECNET or SNA, even though the

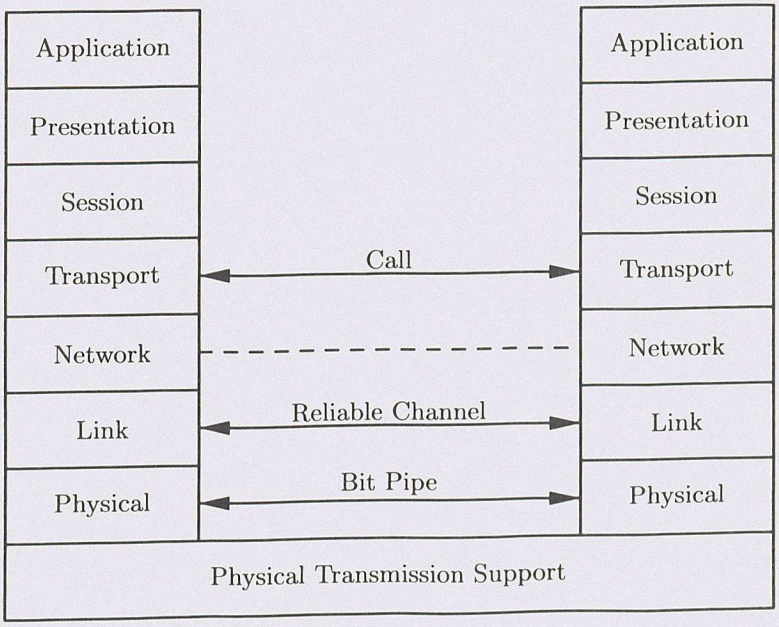

Figure 1.2 OSI Seven-Layer Model

layers of these implementations may not match the layers of the OSI model precisely. The first seven layers of the OSI model are shown on Fig. 1.2. We explain the functions of only the first four levels, however, since they are the levels relevant to this book. More detailed descriptions appear in [1,2,3].

Point-to-Point Layers

The first two layers of the OSI model deal with communication between two points connected by some transmission system. The service offered at the physical layer is the transmission of bits between the two points by the use of electrical signals. This service is implemented in devices such as modems, which are responsible for the electrical transmission of digital information. The nature of this transmission — for instance, whether it is synchronous or asynchronous — is not relevant to this discussion. The physical-layer service is often called a "bit pipe" since this is its appearance to the level-one user.

Nowhere in the level-one service is there any guarantee that data being transmitted in the bit pipe will be received correctly at the other end. This is the preoccupation of the level-two service — the reliable transmission of blocks of data over the pipe — which is done by known methods of error detection and

error correction. These methods are the subject of much of information theory; numerous excellent references exist [4]. In practice, the level-two service adds information to the data being offered by the upper layer that is used by the peer process at the other end to detect and correct errors, if necessary.

Although other functions, such as link flow control, may be added at the level-two layer, the level-two service is succinctly described as a *reliable channel* for the point-to-point transmission of information. This function is the one we consider in the remainder of this discussion.

Transport Layer

The level-three and level-four layers, more complex than the point-to-point layers, are called the network layers for reasons that will become obvious shortly. To clarify the respective roles of layers three and four, we first offer two examples of telecommunication applications and show how they are related to the level-four services. The level-three functions are then presented as intermediate stages between levels four and two. It should be noted that the description here is somewhat simplified; a full description of the network layers appears in [1,2,5,6].

Consider the user of the computer system of a company connected to some local network, say in North America, who wishes to communicate with another computer of the company, say in Europe. The user issues a command to his or her local network requesting the connection and, if it is granted, proceeds with the terminal session exactly as if connected to the local computer. The actual transmission of information and control required to achieve this connection is totally invisible to the user, and is carried out by other lower levels of the OSI model.

Next consider the case of long distance telephone calls. Here, too, the user requests a connection by sending a number of digits to the local telephone office. Again, if the connection is made, the exchange of information proceeds exactly as if the user were directly connected to his or her correspondent, without any knowledge of the intervening stages in the connection or transmission process.

The similarity between these two modes of operation is not coincidental since both are functions of level four of the OSI model. In fact, at this level, the examples are indistinguishable, except perhaps for the fact that the connection request is being made using different signaling techniques: the ASCII (or EBCDIC) code for the first one, and pulse or tone dialing for the second. In both cases, the user is given a "pipe" between two points that can be used to transmit whatever is desired; at this level, the system look like a direct connection between the origin and destination terminals.

With these examples in mind, we can now describe the main function of level four. It provides the user with a connection, sometimes called a *virtual circuit*, from one point to another, that can be used to transmit blocks of

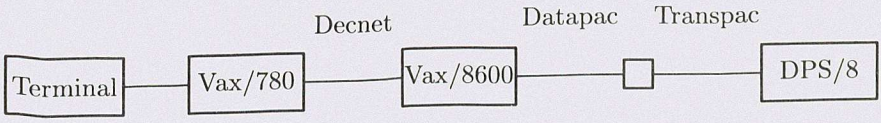

Figure 1.3 A Multisystem Connection to a Remote Host

information. More precisely, the transport layer provides a full duplex channel for the transparent transmission of normal data of unlimited size, as well as priority control data of limited size. This layer also provides independent flow control and a grade of service that can be parametrized. A connectionless version of the transport layer is currently under study in the CCITT but plays no role in the discussion that follows; see [1].

Although conceptually the level-two and level-four functions are similar, there is a crucial difference: In general, in level four, a direct physical transmission system need not exist between the two points. Logically, however, the requirements are the same: There may be delay constraints, maximum permissible error rates, a requirement for maintaining the order of blocks of information, flow control, protection or security requirements, and so on. In fact, many of the level-four techniques for error control and recovery, flow control, and sequencing are identical to those used within level two. Again, this is not surprising in view of the logical similarity between the functions of the two layers.

Network Layer

The transport-layer service provides a virtual circuit that meets the requirements of the user, without any indication as to how this is actually carried out. In principle, the user of a virtual circuit should be unable to decide whether he or she is directly connected to the destination or whether a more complex arrangement is required, and how this is carried out. In practice these more complex arrangements are often required, as in the preceding examples. In the case of this author, the connection to a remote computer in France looks like that of Fig. 1.3: It goes through the local area network, then through the packet-switched network of the country of origin, then the packet network of the destination country, and finally the local network of the destination computer. The same is true for the telephone users, with the national telephone networks at both ends and their local area networks replacing the local area networks of the computer user.

In both cases, the connection is made through a number of telecommunication systems, each of which implements the first three layers of the OSI model (see Fig. 1.4). In order to provide the virtual circuit, a number of physical

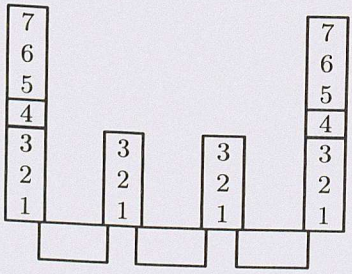

Figure 1.4 Telecommunication Systems in the OSI Model

devices must be made to operate cooperatively. The functions of the level-three layer are those required to implement this cooperative operation of the telecommunication systems. Network-layer protocols, responsible for setting up, maintaining, and clearing connections, provide independence to the functions of the higher layers from the data transmission and switching technology. The network layer includes a number of features worth explaining.

One difference between the network layer and the transport, link, and physical layers is that no single pair of peer processes operates at the end of a virtual connection. Instead, a multiplicity of processes must exchange information within the layer in order to establish and maintain the connection. As a consequence, the network layer is generally distributed geographically in many processes, often one such process per switching point along the connection.

The network layer may have many of the functions of the transport layer already implemented in it. A well-known case is the X.25 recommendation for packet-switched networks, which provides a virtual circuit between a pair of nodes with all the functions of the transport layer already implemented at the network layer. In this case, the distinction between the two layers becomes blurred; in practice, the two levels are merged into one. In fact, the two layers are in many ways redundant; one may ask why they have been defined this way. One plausible explanation, frequently denied by the standards organizations, is that the transport layer was the realm of the data processing organizations, while the network layer belonged to the telecommunication people. Even if not the case, it is still convenient to view the two layers this way because the redundancy becomes more natural (especially when considering the strict separation between data processing and telecommunication organizations when the standard was initiated in the mid-1970s.)

For the purpose of this discussion, we maintain a sharp distinction between the two layers, with the transport layer in charge of the virtual call and the

network layer responsible for setting up, maintaining, and clearing the call through a number of telecommunication systems. Making this separation gives a unified view of switching methods and describes network functions in terms of resource allocation to the calls (see Section 1.2).

Other Layers

We have examined the first four lower levels of the OSI model in the context of an interactive session on a remote computer, or a long distance telephone call. Three additional levels are currently provided for by the OSI architecture. The session layer provides services required for setting up a session: directory assistance, access rights, and so forth. The presentation layer processes information before it enters the virtual circuit, whether for encryption, data compression, or code conversion. Finally, the application layer is defined as the functions that are required beyond those provided by the other six layers. These upper layers are not discussed any further since they play no further role in this book. Again, the interested reader can find more information on this subject in the references; see [1, 2, 3].

Traffic and Layer Coupling

The OSI model has been constructed in such a way that the interprocess communication at a given level is completely isolated and independent from the processes at the other levels. This does not mean, however, that the various layers are not related in other ways. In fact, a major problem of traffic engineering is to study this dependence and to determine to what extent it can be ignored — or how to take it into account when it cannot be ignored. The coupling comes from the particular mode of operation between layers, where a service request at one layer generates one or more requests at the layer immediately below.

Returning to our electronic mail example, the requests for sending messages arrive at random, according to some stochastic process. Suppose that a user requests that the same message be distributed to all the members of a mailing list. Assume that each message is sent by opening a virtual call to the destination. The generation of a mail request in the mail level produces in turn a number of requests for connection in the transport level, below it. Similarly, once a virtual connection is established, it generates a number of requests to the network layer to transmit the information, either in packet form or as an ordinary call. This process repeats downward until the physical level is reached. This increasing cascade of service requests is one of the two forms of coupling between the layers that must be taken into account in conducting a performance analysis of the network.

The second form of coupling operates in the other direction, from the lower to the upper levels. This type of interlayer coupling occurs because some of

the decisions that must be taken at a given level depend on the state of the level immediately below. In our example, a request to distribute a message over a mailing list may be refused, or postponed, if the number of virtual calls currently open exceeds some threshold above which the quality of operation in the computer would be degraded unacceptably. Similarly, a particular request to open a virtual call may be denied if the network level is experiencing a high level of delay or loss. Finally, the path chosen by the network level to connect a call depends to a large extent on the state of congestion of the link level throughout the network.

A major task of traffic engineering is to compute whatever quantities are of interest at a given level, for instance the probability of blocking for a call, from the quantities available at the lower level — say the probability that the individual links in the network are fully occupied when a call arrives. For circuit switching, the interrelation between the transport and link layers is fundamentally important, and will be the subject of a large part of this book. From the brief description just given, we can see the types of disciplines that are relevant to traffic engineering. Because of the statistical nature of the requests at each level, the methods of probability theory must be used for the analysis. Also, because of the coupling between the levels, decomposition methods and their counterparts, iterative solution techniques, also play a fundamental role. Optimization techniques are also used because most of the problems relate to the synthesis of systems.

1.2 Switching Methods

The description of network operation in terms of the OSI model is useful to classify the various functions that are required in this process. The OSI model also provides a synthetic view of the system, which in turn can be used to relate some aspects of networks that were previously seen as distinct. A case in point is the so-called switching methods used in telecommunication networks today — which, as we now see, are not really switching methods, but rather different types of processes for allocating resources to calls.

Current telecommunication networks fall into two very general classes: circuit switched and packet switched. These classes are often presented as totally different, and their modes of operation distinguished in such a way as to make it difficult to take a synthetic view of the two switching methods. Here we show that, within the context of the OSI model, packet switching and circuit switching share some underlying features, at least from the point of view of the transport and network layers, but also have some essential differences. In addition to giving a clear picture of the similarities of and differences between the two switching techniques, this unification provides a framework in which to

classify the new switching methods that are currently being proposed for new services (see Chapter 10).

We have seen that the network functions are responsible for connecting and maintaining calls between two processes operating in the transport layer that is within and between the telecommunication systems used to carry the information from one end of the network to the other. In other words, when a call request is set up, the network makes a commitment to have some resources available for use by the call if and when they are required. In this sense, the network layer protocols are responsible for allocating transmission resources to the call. The particular way in which this allocation is made is the fundamental difference between the switching methods, and is determined to a large extent by the characteristics of the call.

In our discussion of the OSI models we gave two examples of calls: (1) an interactive session on a remote computer and (2) a long distance telephone call. Other examples are a file dump from a computer on a line printer, a teleconference, and the transmission of a still picture. In general, the flow of information within the call is not constant, but comes in bursts of various lengths. For instance, for the interactive computer user, there will be short bursts corresponding to the typing of commands, rapidly followed by longer responses from the computer, then followed by quiet intervals corresponding to the "think time" of the user. In this case, the utilization, defined as the ratio of busy time to the total duration, is often quite low. On the other hand, calls corresponding to a file transfer have a short burst corresponding to the request followed by a long continuous stream of activity; the utilization will be quite high. Finally, telephone calls have a utilization that falls in between, typically around 50%, and other types can have utilizations that range from zero to one. The utilization factor is one of many parameters used to describe a call. Others are the origin and destination, bandwidth (average, peak, or other), delay constraint, and admissible error rate. The common features of switching techniques come from the realization that whatever the method, the same three decisions must be made; these decisions are arrived at in the same way, based on the call parameters and the current knowledge of the network state.

The first decision to be made is access control, that is, whether to accept the call or not. Then, if the call is accepted, what telecommunication systems should be used to connect it must be determined. This is the resource-allocation process mentioned before, and it is here that the two switching methods differ. Note that these two stages are interrelated since the decision whether to accept a call may very well depend on the path used to carry the call if it is accepted. This, as we said before, is what provides the coupling between the two levels. Finally, flow control is the process that ensures that the call parameters remain within the specified values. This process is relevant only for virtual circuit switching, where the connection request may specify a maximum instantaneous

bandwidth that the call will use during its duration. Once the call is set up, the network must make sure that the call does not exceed this threshold by sending a large burst of packets in a very short time. This is particularly important for new services, such as video, where the process by which packets are generated within a virtual call is highly bursty. We now discuss how the first two stages are carried out for the two switching methods, indicating how flow-control issues may arise in these systems.

Circuit Switching

Every time a connection request is made, the requirements in terms of bandwidth, duration, utilization, and so on are known. A simple allocation strategy is to reserve immediately all the resources required in all the telecommunication systems that link the origin to the destination, and to keep them reserved for the complete duration of the connection. This is the essence of circuit switching. The task of the network-layer processes is first to determine whether such an allocation is possible. The decision rule may be that the call is accepted whenever there exists a path in the network with sufficient bandwidth to meet the call requirement, or it may be based on more complex measures of the future impact of accepting this call at the present time. If the call cannot be accepted, it is said to be blocked, and either can be made to wait until the resources become available, or can be cleared. If the call is accepted, the network processes must make sure that the reservation of capacity is correctly carried out along the selected path. Here again, the routing rules may be quite simple, such as to take the first path available from a fixed list, or may depend on a more complex measure of the network state, such as taking the path with the largest expected residual capacity. After this connection stage, the network processes have very little work to do while the call is in progress except to ensure that capacity is correctly released when the call terminates. Because of the complete reservation inherent in circuit switching, the transmission systems are already available whenever the call has something to transmit.

Circuit switching has traditionally been used for telephony, where a physical connection between the two ends of a call was established using a real circuit through mechanical relays. With the advent of frequency and time division multiplexing, this circuit is now replaced by a frequency band or a time slot. Conceptually, it is the same as traditional circuit switching in the sense that the particular band or time slot is not available to anyone else as long as the call remains connected.

Because there is an allocation of physical resources, circuit switching is well suited to calls with a high utilization, such as file transfer or uncoded voice, as well as to calls that require little or no delay in the transmission of information, such as voice or real-time applications. The only delay is caused by propagation in the medium, generally negligible except perhaps when satellite

links are involved. Also, because of the minimal amount of work necessary to maintain the connection, a large number of calls can be processed through a given switch, as in large public telephone networks. Finally, circuit switching provides complete isolation between customers after the calls are set up, making flow control within the call unnecessary.

Circuit switching as it currently exists also has a few disadvantages, some more important than others. Most frequently mentioned is the impossibility of tailoring the bandwidth allocated to the instantaneous requirements of the call and the ensuing poor utilization of bandwidth for some types of applications. The importance of this defect depends on the relative costs of bandwidth and switching; it is rapidly becoming less relevant with the advent of very large-capacity optical transmission systems. More important are the long set-up times, of the order of seconds or tens of seconds, which prevent the use of circuit switching for applications with stringent real-time constraints.

Another defect of circuit switching appears when considering calls with large bandwidth requirements. Current switching systems were designed primarily for voice communications, and can switch only calls having the bandwidth of a single voice call. As a consequence, wide-band calls cannot be switched as a single entity, but must be demultiplexed into individual voice calls, switched, and then remultiplexed into the wide-band call. This procedure is quite complex and requires additional multiplexing in the switches, making the use of circuit switching for wide-band applications less attractive. These last two defects cannot easily be eliminated with current switching technology. Preliminary work on fast circuit switching and bulk switching in cross-connect switches, however, indicates that these capabilities are feasible and could become available in the next generation of switches.

Packet Switching

For calls with a low utilization factor, circuit switching is inefficient — the allocated resources are used only a small fraction of the time. Packet switching is designed specifically for such calls. Currently packet switching is used primarily for computer communications. The call is generally accepted when it arrives, although in many networks the call can be rejected if the network is congested. The difference from circuit switching is that no physical resources are allocated until there is actually something to transmit. When there is, the information to be transmitted is divided into subunits, called *packets* or *messages*, and network resources are allocated for individual packets, and only for the duration of the transmission from one switching point to the next.

Here the network processes must first determine whether the call should be accepted, as with ordinary circuit switching. The decision rule may be a threshold for the expected transit delay across the network at the time the call request is made, or may depend on other parameters, such as average or peak

bandwidth. Then the routing decision must be made for accepted calls. Again, this could be based on an estimate of the network state, on characteristics of the available paths, and on the call parameters. Typically, the call is connected on the path with the shortest expected end-to-end delay within a given list of available paths.

So far, this procedure is conceptually identical to the access control and routing for circuit switching, the only difference being the particular choice of call parameters and network states used to reach a decision. The main difference between the two modes is that, at this stage in packet switching, *no physical resources are allocated to the call*. This allocation is made every time there is something to transmit within the call, and only on a point-to-point basis, that is, at the link level. Thus the work to be done by the level-three processes during the duration of the call is significantly greater than in the case of circuit switching. On the other hand, this procedure is clearly more efficient than circuit switching in terms of bandwidth utilization. Also, flow control within the call must be provided to ensure that a user does not exceed his or her allocated bandwidth or some other parameter associated with the call.

A Unified View of Switching

We can now express more clearly the difference between circuit and packet switching as viewed from the network and transport layers. In circuit switching, a network-wide resource allocation is made for the complete duration of the call (i.e., at the transport level), while in packet switching, a step-by-step allocation is made as packets progress within the network on a need basis (i.e., at the link level). For circuit switching, there is a real commitment of resources at call connection, while for packet switching, this commitment is only virtual. Put another way, the allocation of resources to a circuit-switched call is deterministic; to a packet-switched call, it is stochastic. Nevertheless, even with this probabilistic allocation, the process of call acceptance and routing for packet-switched calls is very similar to the corresponding processes for circuit-switched calls: The statistical effect of the allocation can be calculated, at least in principle, whenever the characteristics of the call and the network state are known. The result can be used for access control and routing decisions based on the consequences of this virtual allocation of resources to the call and its impact on the performance of the network, just as with circuit switching.

For example, from the knowledge that the call request is for an interactive session on a remote computer, it is possible to compute the average bandwidth required for this call on some path in the network. Accepting this call means that the network must be prepared to allocate this much bandwidth *on the average* to this call on the path selected to route the call. This in turn increases the expected delay on this and other network paths, information that can be used to decide whether to accept the call and where to route it. Average delay

is only one of many parameters that can be used. Other measures of network utilization are available, such as a percentile on the delay distribution, although most such measures are more difficult to compute than the average.

This decision process is conceptually identical to the connection process for ordinary circuit switching, except that the call and network state parameters are different and that expected instead of deterministic values are used. The unifying view comes from the fact that even though one method uses a real allocation of resources while the other uses only a virtual one, the call-connection process as viewed from the transport layer is fundamentally the same, and can in principle be described very similarly for both switching techniques. For this reason, much of what will be said in this book should apply equally well to the routing of virtual calls and to traditional circuit switching.

Finally, note that the policy used in the presence of call blocking is in no way a characteristic of the switching method. Although it is customary to use a blocked-calls-lost policy for circuit switching, and a blocked-packets-delayed policy for packet switching, this distinction is irrelevant to the switching method used. Current telephone switching machines can make blocked calls wait, a process known as *call queuing*. The other possibility, that of packet switching with blocked packets lost, is a perfectly legitimate mode of operation for information that can tolerate loss, such as voice, although apparently this method has not been used much. A notable exception is the knockout switch, where a small packet loss probability is tolerated in return for a large increase in throughput [7, 8].

Scope of Circuit-Switching Methods

As its title indicates, this book is about circuit switching, that is, about levels four and three of the OSI architecture. Most of the theory expressed here comes from the area of telephone network operation and planning, which use ordinary circuit switching. This may seem to limit the scope of the techniques discussed to telephony. We expect, however, that circuit switching will be used for quite some time for voice, as well as for new networks. In the first case, the large base of installed equipment used for telephone service will remain in place for at least the medium term; the shift to new integrated networks must be evolutionary. In the longer term, it is not clear whether a synchronous mode of transfer such as circuit switching or an asynchronous mode of the packet-switching type will be used in the integrated transport network.

Another possibility, which seems increasingly attractive, is a unique hybrid switch, capable of both circuit and packet switching. Circuit-switching methods would continue to be used for a significant fraction of the traffic. There is also an emerging consensus that if the new networks are of the packet-switching type, this will be done in a connection mode. As the discussion in this chapter indicates, many of the results of classical and more recent circuit-switching

methods will become relevant to these virtual circuit-switched networks. As shown in the discussion of the OSI model in Section 1.1, the methods of circuit switching can be applied much more widely than merely for telephone service; they apply to the transport layer of *any* network operating in a connection mode. The reasons should be obvious by now: The access control and the routing decision are the same in both cases; they depend both on the parameters of the call and on the state of the network.

For circuit switching, a call is blocked if the resources are not available at the time of the request. For packet switching, the call is rejected if accepting the call is *expected* to degrade the network performance beyond a certain value. This decision also depends on the call parameters and the network state. Also, the routing of an accepted call can be made to depend on the call parameters and network state in a similar way for both methods. The decision rule is the same in nature; only the call and network parameters used to reach the decision are different.

As of today, this integrated view of access control and routing has not been used in planning and analyzing telecommunication networks. Telephone networks and data networks have evolved separately, using routing and control techniques that are on the surface quite distinct. With the advent of integrated transport networks, a wide variety of virtual calls will have to be served by the same network. Some of these calls will have characteristics that make them suitable for packet switching, while others will be ideally suited to circuit switching. In this case, more complex admission-control and routing techniques for virtual calls than the ones currently used may be required, independently of the actual policy used for allocating resources by the network layer. A unified view of access control and routing through the generalization of the results of classical circuit switching would become necessary; this is the subject of current research. Let us now examine another aspect of this work and its relation to the planning process for telecommunication networks.

1.3 The Planning Process

In this section, we give a short description of the traditional view of the planning process of telecommunication networks and its relation with the contents of this book. As with any decision process, to arrive at its results network planning relies on external information. In case of telecommunications, the forecast of the demand for services over some horizon drives the evolution of the network. Some economic information concerning the cost structure of the elements making up the network is also required. There exist numerous methods of forecasting and engineering economics [9,10] to obtain these values, which are not specific to the field of telecommunications and thus are not discussed here. Another requirement is some knowledge about the technical capabilities

of the available systems, which may limit the options available for the design. Although this aspect is more specific to the planning of technical systems, we again assume that it is given as input to the planning process; it is therefore not the subject of decision.

The planning problem can now be stated as follows: to implement the first four layers of the OSI model and to provide the required physical support. This must be done in such a way as to keep track of the evolution of the demand for services, while also ensuring that a satisfactory grade of service is maintained at all times. The knowledge required for this task spans a variety of scientific disciplines. At one extreme, the definition and implementation of peer processes and protocols is an area of computer science. At the other extreme, the design and implementation of point-to-point transmission systems is a subject of electrical engineering and physics. Network planning as we understand it in this book is the intermediate stage in which these elements are connected in such a way that the OSI model becomes operational.

Assuming that all the protocol issues have been settled and that the transmission technology is known, what remains is a complex, distributed, and dynamic capacity-augmentation problem. Because of its complexity, the problem cannot be solved in a single stage — the traditional, and only feasible, approach is through decomposition. We identify four interrelated stages in the planning process, each with an increasing level of detail in cost and structure.

The first stage is the design of a topological structure for the network: where to place the components and how to interconnect them, subject to connectivity constraints. Here we use the methods of topological optimization and graph theory. At this stage, the cost information is simplest. All information about the transmission network is summarized into a fixed interconnection cost per unit length between offices. Typically, this is a single value calculated from the capacity of the transmission systems actually installed in the network; the actual cost is proportional to the distance between offices. Similarly, the switch costs are given and can take only a limited set of values, one for each switching technology available; these costs do not depend on the traffic volume through the switches. The output of this step is a connectivity matrix and, in some cases, the optimal location of switches or concentrators.

The network-synthesis problem uses this information to calculate the optimal size of components, that is, the transmission and switching systems, within the topology specified by the result of the topological optimization stage, and subject to grade-of-service constraints on network-performance measures such as transit delay or loss probability. A linear cost model and fractional sizes for the network components are used. Because the constraints are not linear, the methods of nonlinear optimization are most frequently used at this stage. The network synthesis is made up of two distinct and interrelated subproblems: traffic routing and dimensioning. The routing problem is to determine how to connect calls as they arrive, given the topology and size of the transmission and

switching equipments. Conversely, the dimensioning problem is to determine these sizes given a particular routing method, taking into account some constraints on the grade of service (GOS) that must be met by the network. The output of the synthesis stage is a route plan and a set of logical links between nodes, that is, the requirements for transmission facilities between switching points.

The logical links express a requirement for transmission equipment between two points but do not specify the particular technology used to implement the requirement. This is done in the network-realization stage, often called the circuit-routing stage, or sometimes simply the routing stage. Network realization determines how to implement the capacity requirement computed by the synthesis stage using the available components such as a cable of copper wires, a frequency division multiplexed coaxial cable, or a digital transmission system. Each component has a different cost/capacity characteristic function and comes in different modular sizes. The circuits in the copper and coaxial cable can be allocated one unit at a time, while those in the digital system must be used in large blocks, typically of 24 or 30 circuits at a time. Also, the marginal cost of these systems may be different — for instance, because of the A/D conversion equipment that may be required in some cases, depending on the type of switch to which the transmission system is connected. Similar situations hold for the start-up cost of a new system and for the incremental cost of providing a new capacity module. The usual solution is *not* to route all the demand on the cheapest transmission technology between two points. Systems are subject to failures, microwave systems can be sensitive to weather conditions, and people have the bad habit of digging out telephone cables while in the process of construction. For reasons of reliability, a circuit demand should not be routed on a single system, and conditions should be placed on the solution to force multipath routing in such a way that a minimal grade of service is maintained in case of failures. All these elements are taken into account in the circuit-routing stage. The problems to be solved are generally of the multicommodity flow type, with modular and nonlinear cost functions and with reliability constraints.

Obviously, these four stages are interrelated and the planning process is iterative, embedded with many levels of iteration (see Fig. 1.5). The output of the circuit-routing stage yields detailed information on the actual transmission cost between two switches. Given the current size of the logical link, a new value can be computed for the average cost per circuit, which can be used as input to another synthesis stage. Similarly, after this synthesis-routing process has converged, new values for the interconnection costs are computed, and a new topological optimization step can be performed.

In practice, only a few iterations are required before a satisfactory solution is found, and not all steps of this process need be performed each time. The topological structure of a network is not changed very often, especially that of

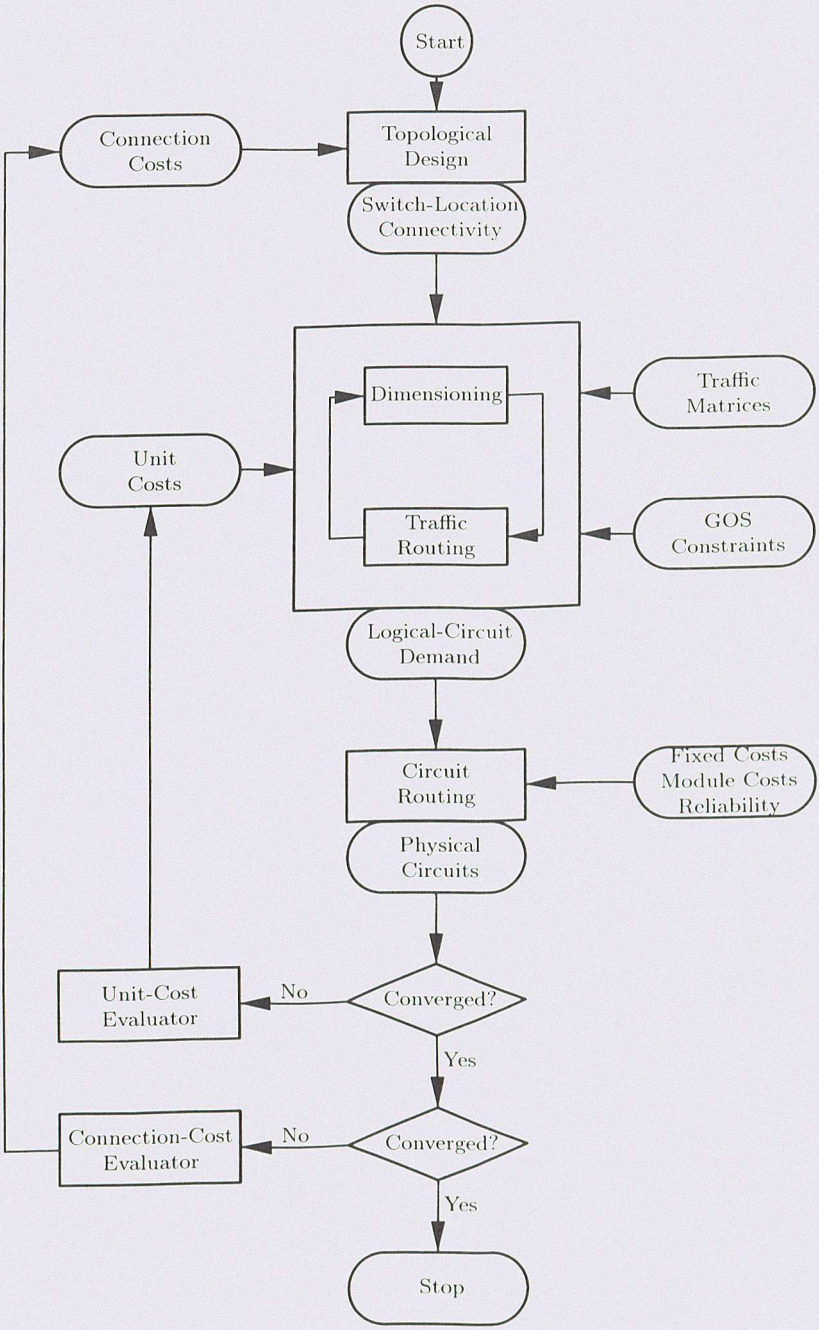

Figure 1.5 The Planning Process

a long-established network such as the telephone network. Similarly, the *type* of routing used in the network is not changed often — typically once every 20 or 30 years. The component sizes, such as trunk group capacity, change more frequently, typically every year and sometimes more often in the case of demand servicing. This time scale is likely to become smaller with the introduction of dynamically rearrangeable networks using slow switches, which allow a network to be reconfigured on a demand basis. Finally, on a shorter time scale still, the parameters of a particular routing method may change on an hourly basis, and, in the case of fast adaptive methods, on a call-by-call time scale in some cases.

A dynamic element is often required in the planning process. In such a case, the evolution of the demand is predicted for a number of years, and an evolution strategy must be mapped over this horizon. The process is still represented by the diagram of Fig. 1.5, but with an explicit dependence on time. Although the problem is conceptually not very different, the numerical difficulty is much greater; as a consequence, the planning process may be simplified even more.

The topological design and circuit-routing stages arise in many areas besides telecommunication networks, such as transportation and power networks. There is an abundant literature on this subject, most of which applies to the telecommunication network problem [11,12,13]. Because of the traffic-routing component, only the network-synthesis stage is specific to telecommunications, and the particular models and techniques used to solve it depend to a large extent on the switching method. The synthesis problem for packet-switched networks has been amply studied and documented in the open literature; there are many good references on the subject [2,3,14].

The synthesis problem has also been solved very efficiently for conventional circuit-switched networks, but, as stated earlier, this information is scattered in a number of conference proceedings and is generally not as easily accessible. Therefore we provide a more synthetic view in this book.

1.4 Notation

We are now ready to begin an in-depth study of circuit switching. First, however, there are a few points about the notation used throughout the book. Of principal importance is the distinction between variables representing quantities defined across a network (e.g., the average time for a message to go from its source to its destination node) from quantities that are defined for the two ends of a link (e.g., the average time for a message to be transmitted from point a to point b over some transmission system).

As a rule, superscripts such as i, j, and k denote end-to-end variables, while subscripts such as s and t refer to variables related to a particular link. We use equivalently a two-or one-index notation for double-index variables. It

is assumed that there is a mapping $o(k), d(k)$ that gives the number of the origin and destination nodes o and d for a given value of the single index k, and conversely that there is a mapping $k(o, d)$ that yields the index for any origin-destination pair. With this notation, the probability that a telephone call from an office i to an office j cannot be connected is represented by $L^{i,j}$, or equivalently, by L^k, where $k = k(i, j)$. This is defined even when no direct link exists between nodes i and j. The probability that a particular link s of the network is blocked when a call attempts to use it is given by B_s, or equivalently by $B_{m,n}$. This convention is used consistently throughout the book.

Other indices besides those denoting the origin and destination are introduced in some cases. This is done in such a way that the use of subscripts and superscripts follows this convention whenever the distinction is meaningful. In other cases, the precise meaning of the index should be clear from the context. A detailed list of the symbols used in this book can be found in Appendix C.

References

[1] Stallings, W., *Data and Computer Communications*, Macmillan, 1985.

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

[3] Bertsekas, D., and Gallager, R.G., *Data Networks*, Addison-Wesley, 1987.

[4] Gallager, R.G., *Information Theory and Reliable Communication*, Wiley, 1968.

[5] Green, P.E., *Computer Network Architectures and Protocols*, Plenum, 1982.

[6] Green, P.E., "Protocol conversion," *IEEE Transactions on Communications*, vol. 34, pp. 257–268, 1986.

[7] Eng, K.Y., Hluchyj, M.G., and Yeh, Y.S., "A knockout switch for variable-length packets," *International Conference on Communications*, pp. 794–799, 1987.

[8] Yeh, Y.S., Hluchyj, M.G., and Acampora, A.S., "The knockout switch: a simple, modular architecture for high-performance packet switching," *IEEE Journal on Selected Areas in Communications*, vol. SAC-5, pp. 1274–1283, 1987.

[9] Smith, R.L., "Optimal expansion policies for the deterministic capacity problem," *The Engineering Economist*, vol. 25, No. 3, pp. 149–160, 1980.