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

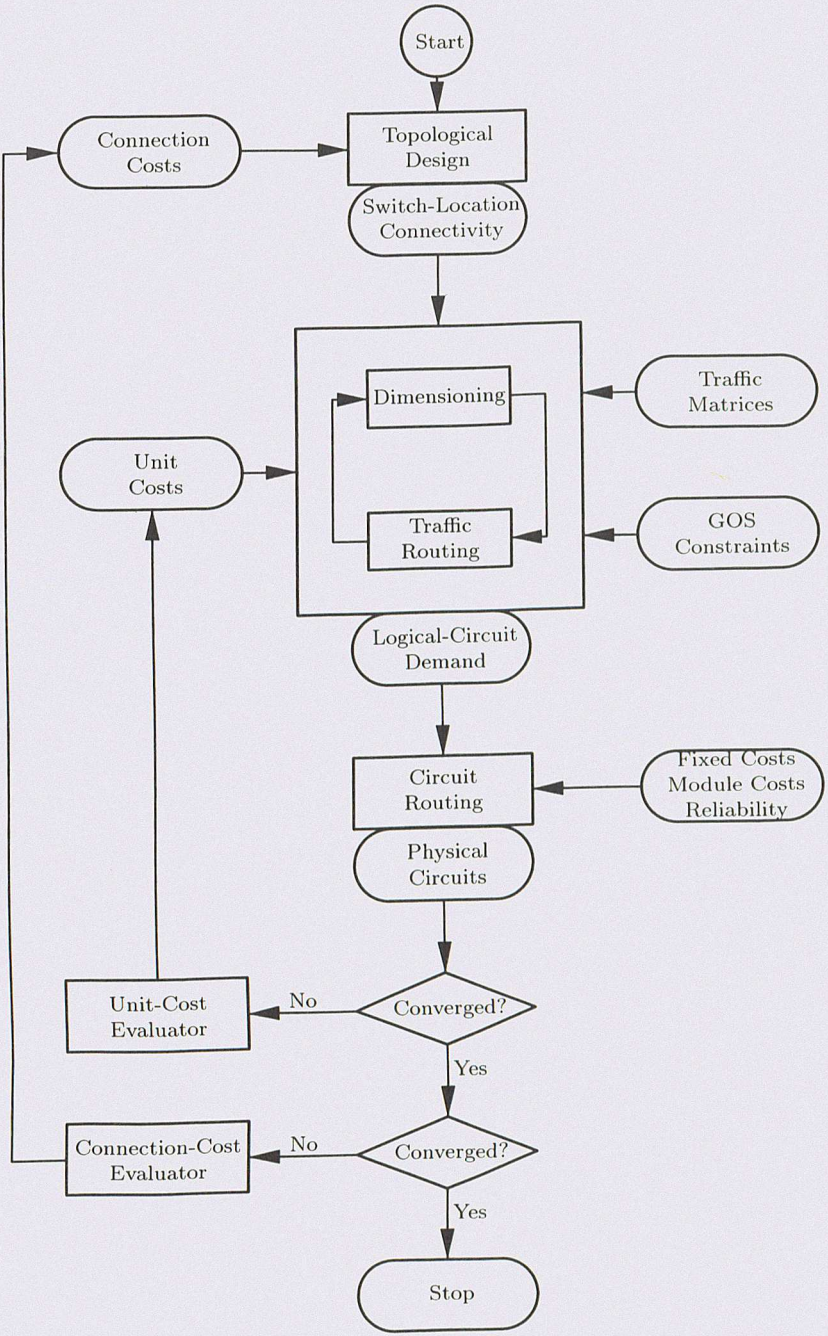

Figure 1.5 The Planning Process

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.

[10] Luss, H., "Operations research and capacity expansion problems: a survey," *Operations Research*, vol. 30, pp. 907–947, 1982.

[11] Frank, H., Frisch, I.T., Van Slyke, R., and Chou, W.S., "Optimal design of centralized computer networks," *Networks*, vol. 1, pp. 43–57, 1971.

[12] Yaged, B., "Minimum cost routing for static network models," *Networks*, vol. 1, pp. 139–172, 1971.

[13] Yaged, B., "Minimum cost routing for dynamic network models," *Networks*, vol. 3, pp. 193–224, 1973.

[14] Hayes, J.F., *Modeling and Analysis of Computer Communications Networks*, Plenum, 1984.

CHAPTER 2

Routing Techniques and Models

The performance of a network depends on such factors as the network configuration, the offered load, and the network management methods. An important element of network management, called *network routing*, consists of the decision rules used to connect the calls as they arrive at the network; a variety of methods are now possible. We describe in depth some of the methods that are currently in use or proposed for implementation in large circuit-switched networks. We also introduce useful graphical models to represent these methods, and briefly cover some mathematical models that can describe them. The question of evaluating the performance of a particular network operating under one of these algorithms is discussed in Chapter 4; the question of choosing the optimal routing method, in Chapter 7.

2.1 Definition of Routing

Consider the real-time operation of a network. Whenever a new call arrives at an office, we must determine whether there exists a path on which the call can be connected to its destination. If a path is found, we must decide whether to use it or not; if there is no available path, we must decide what to do with the call. Each of these steps is related to a different aspect of routing.

The selection of an available path constitutes the routing problem in the strict sense, requiring the paths available for each stream to be defined — an important part of the specification of a routing technique. This definition could be the set of all paths permitted by the network structure or, more simply, some convenient subset. It must also include a description of a procedure for finding such a path and for selecting which path to use if more than one is available.

The decision of whether or not to connect a call on an available path is often called *flow control*. Flow control can be viewed as a special case of routing, where lost calls are routed on a fictitious path of infinite capacity. Although the flow-control problem has been extensively studied for packet networks, no comparable work exists for circuit networks. In this book we do not consider the flow-control problem, but rather assume that a call that can be connected always will be connected.