Planning and design of data communications networks

by WUSHOW CHOU

Network Analysis Corporation
Glen Cove, New York

INTRODUCTION

The worldwide trend in computers in the 1970's has been the movement from off-line to on-line systems and the growing integration of computing and communications. It is expected that 70 percent of all computers will be connected to terminals via communication facilities by 1980. This trend has stimulated new applications for integrated systems, generically termed data or computer communications systems, and has broadened the current user base for existing applications.

As a result of this, communications systems are becoming increasingly larger and complex. In addition, innovations in system design concepts, hardware features and transmission services have accelerated. It is not easy for a communications network designer to cope with the complexity and the innovations that are being continually introduced. When he is just becoming accustomed to character oriented data communication control procedures as a way of life, he is told that the advanced control procedures are bit-oriented. When he is just about to understand the concept of store-and-forward packet-switching for computer communications he is exposed to "ring" structured network concepts. When he has begun feeling comfortable with multiplexers, he is faced with programmable concentrators and front-end processors with endless options. When he is just about to think that modems and analog signals are natural ways to transmit digital information, he realizes there will be digital transmission services and digital interface units. When he is still struggling to understand and distinguish among the various terrestrial transmission services offered by common carriers, he is confronted by the emergence of domestic satellite communication services that may save him money. The list of these dilemmas is endless. A system composed of several hundred terminals used to be classified as "large"; now there are systems with thousands of terminals. Many existing systems are experiencing a rapid growth in traffic volume, and a corresponding increase in response time. It is fair to say that almost every data communications analyst and network designer has his share of worrying and uncertainty about his system. Rules of thumb obtained from previous practical experience alone are no longer adequate except for very simple cases. Or, for some situations, there is no such rule at all.

How should a manager conduct the planning or upgrading of a modern data communication network? How can a designer or an analyst evaluate the best system concept and design strategy for his system? How can a user select the best cost/performance communication devices and transmission services for his system? What are the impacts of new technologies and the proliferation of new devices, new services, new tariffs, and new vendors? Are there design and analysis tools available to assist the designer or analyst make his decision?

The purpose of this paper is to present concise answers to the questions posed. These answers are an overview of the state-of-the-art techniques and approaches in the planning and design of data communications networks. Rules of thumb and simple concepts that are readily available in trade magazines are not emphasized. Instead, we will concentrate on the new, sophisticated techniques which have been specially developed to handle the complex problems present in modern data communications networks, and are used by leading experts in the field.

Some of the above mentioned questions are treated in detail in Lynn Hopewell's, "Planning in the Data Communications Environment," Pat McGregor's, "Effective Use of Data Communications Hardware," Mario Gerla's, "New Line Tariffs and Their Impact on Network Design," and Aaron Kershenbaum's, "Tools for Planning and Designing Data Communications Networks." Another purpose of this paper is to provide linkage and continuity among these four papers. Terminology or concepts that are mentioned in this paper but not discussed in detail in this paper or any of the above four papers is likely to be found in James Martin's, System Analysis for Data Transmission.

MANAGEMENT AND PLANNING

In most large communications networks currently in operation, costs can easily be reduced by 15 percent or more with only minor alterations in the network. The cost savings can often reach 30 percent or more if reoptimization of the whole network is allowed. Furthermore, the cost reductions can be realized without degradation in performance. Indeed, both cost and performance can often be improved simultaneously.

The implication of the above statement is that the existing
networks are not properly planned and designed. Why is this so? If one has to point a finger, the traditional corporate structure is to blame.

Traditionally, corporate communications systems have been left in the hands of middle or low level managers, mostly with limited technical background. With the emergence of data communications, which demands a broad range of knowledge in both practical engineering and advanced theory, the top management finds it even more undesirable to exercise direct control over the planning and operation of communications networks. Let us now examine the consequences of this lack of interest and understanding of the complexity of the problem.

Middle level technical managers who often have the major responsibility for planning and operation face many difficulties. First, these network managers are usually technically oriented. They often do not have enough training in management. They are also likely to have the characteristics common to almost all technical people, overestimating their ability and overcommitting themselves for technical projects.

Making the whole planning and operation process worse is the fact that their hands are usually tied. It may be relatively easy to get approval from the top management for tangible expenses such as buying communications equipment and paying line costs. It is considerably harder, however, to obtain funds for less tangible expenses, such as program development and seeking expert assistance during the planning and design process. Thus, they must make compromises within the limitations of their budgets. Furthermore, they do not want to risk user dissatisfaction. As a result, the tendency is to overdesign networks, using more lines and equipment than is actually necessary. So the small saving in the planning stage often results in a large increase in the cost of the final network. Also, the overdesign may not guarantee satisfactory performance.

Due to their relatively low level in the organization hierarchy, the network managers do not want to chance any possible blunder. This usually means that they would hesitate to initiate any innovation or change unless it is absolutely necessary.

Another difficulty is the lack of coordination among different groups providing technical assistance to one another in a broad data processing system or a larger corporate communications system, of which the data communications network is a part.

Probably the single most significant factor that contributes to the inefficient designs is the general underestimation of the complexity of data communications. Data communications networking is an evolving, new technical field. To be able to master this field, one needs to know computer software and hardware, communications hardware, transmission facilities, line tariffs, human psychology, queueing theory, statistics theory, communications theory, advanced computational techniques, and the most advanced network optimization techniques. Some of the advanced techniques may be the result of the most current research. He must be aware of the new developments in the field. In addition, he must be clever and think innovatively. There are not many people who possess all these capabilities. Without knowing this situation, a big corporation's management is likely to think that somebody in the company should be able to develop an expertise in the area of data communications networks, or at least it should be easy to hire someone with this expertise.

In some cases, usually out of desperation, assistance from consultant is sought. However, with the notion that any of the many data communications consultants can do the job, technical competence does not always play enough of a role in the selection of a consultant. A consequence of this is that one may not get the necessary expertise.

This whole section has been discussing the negative part of the management planning process. What is the positive side? In this author's viewpoint, but not the author's alone, the improvement can be achieved by upgrading communications management to the corporate management level. The network manager with management background should force himself to be technical; the network manager with technical background should force himself to acquire management ability. There is actually a trend, which is taking hold, though slowly, toward the upgrading of the communications management and the merging of both technical and managerial talents into the management. One can observe this trend by the fact that the chief executive officers of some of the new specialized common carriers have reputations of superb ability in both the technical field and management.

REQUIREMENT ANALYSIS AND DESIGN CONSTRAINTS

Successful implementation of a data communications system depends to a great extent on the thoroughness of the data traffic analysis and the user requirement analysis.

Gathering traffic information is the most tedious part of the planning. Traffic information should be collected from every user by means of current measurements and future projections. To be included in the traffic information for each terminal and for each type of message or transaction are distributions of the number of transactions per unit time during the peak hour, average day and peak day, input message length distribution (number of characters per message), output message length distribution, and priority. It is almost impossible to accurately measure and project this information. However, one should try his best. It is often helpful if the planner visits the users to assist them and to validate the information supplied by them.

Even more uncertain than the traffic statistics and projections are the users' requirements on network performance. In general, users do not know exactly what they want. Sometimes they may demand a level performance which is practically impossible. Other times, they demand a performance level that they may not need but must pay a high price to attain. If cost is not a concern, any user would like to see negligible response time and almost perfect network reliability. Apparently, no one can afford to pay for such level of performance. But users usually do not have the feeling for the cost/performance relationship. It is the
planner's and the designer's responsibility to educate the
users, to show them the relationship between cost and per­
formance, and thereby assist them in modifying their per­
formance requirements to reasonably obtainable levels. The
traffic information and the users' performance requirements
thus form the constraints for the design.

PERFORMANCE CRITERIA AND
CONSIDERATIONS

A general goal in designing a data communications network
is "to design a minimum cost network satisfying performance
requirements or criteria." A common slogan in this field is
"improving the cost/performance ratio." What is performance?
It means different things to different people. For a
well designed system, the performance should be measured
by the following criteria: blocking probability and/or message
response time, traffic capacity or throughput, network reli­
bility, transmission error rate, and sensitivity to variations
in traffic level. Knowing the traffic bottleneck is also an
important piece of information, especially for future expan­
sion and upgrading. However, it is not generally used to
measure current network performance.

Blocking probability

This criterion is used to measure the promptness with
which a data communications system responds to calls from
dial-up terminals. It can be defined by any of the following
three factors:

1. At least "A" percent of calls obtain access to a com­
puter port within "B" minutes (this definition is used
for systems in which to have a connection, the terminal
must first dial to a switchboard; an operator then
attempts to connect it to an empty computer port).
2. At least "C" percent of calls obtain access to an
empty port on the first attempt.
3. "D" percent of calls obtain access to an empty port
with no more than two attempts, three attempts, etc.

The parameters A, B, C, and D are constants determined
by user requirements or network planners. For example,"A" can be 99, "B" 5, "C" 95, and "D" 99.

Message response time

This criterion is used to measure the promptness with
which a system responds to terminals connected to the
system by leased or private lines. Message response times
have different definitions at different parts of a data com­
munication system. So far as the users are concerned, "termi­
nal response time" and "overall response time" are most
meaningful. The terminal response time is defined to be the
time required from the instant the "transmit" or equivalent
key on a terminal keyboard is depressed to the moment the
reply message begins to appear at the terminal. This is the
most commonly used criterion. However, it has serious draw­
backs when used as a measure of the "service promptness,"
since a user may have waited a long time at the terminal
before his message is keyed in. The overall response time is
the elapsed time from the instant that a user or a message
arrives at a terminal to the moment the user is completely
served or the reply to that message is received. The response
time requirement is usually defined as "average response
time should be no more than X seconds" and/or "response
time for at least Y percent of transactions should be no
more than Z seconds."

System capacity or throughput

Capacity in its most liberal interpretation is often taken
... to mean the maximum amount of traffic, in terms of trans­
actions per second, characters per second, etc., that a system
can carry. Unfortunately, such a definition would be totally
unrealistic since a user, attempting to send messages into a
system operating at this capacity, would experience intoler­
able long delays from the time he inputs his message to
the time he receives a reply.

For a more practical definition, the capacity is defined as
the maximum traffic that a system can carry, while satisfying
the blocking probability criteria and/or response time re­
quirements.

Network reliability

While the failure rates, MTTF, and MTTR of the equip­
ment and lines are often beyond the control of network
planners, the network's reliability can usually be strengthened
with proper network structures. Again, the definition of
reliability varies according to usage. For most purposes, at
least one of the following definitions is applicable.

1. Percentage of time a terminal can communicate with
the central computer.
2. Percentage of time an office can communicate with
the central computer (the office may have more than
one terminal).
3. Percentage of time a terminal can communicate with
any other terminal (this definition applies when there
is direct inter-terminal communication).
4. Percentage of time an office can communicate with
any other office.
5. Average number of terminals or offices that are con­
nected to the network.
6. Average number of equipment failures, or the average
man hours required for repair, in a day or other unit
of time (this definition is usable for equipment main­
tenance crews).

Sensitivity

A system may behave properly if the traffic volume is
within the projected range, but break down entirely if the
traffic exceeds the volume for which the system was designed. Thus, a good planner should be concerned with the effects that the system would experience if the actual traffic is above the projection. He should make sure that a small variation in the projection does not create intolerable response times, or blocking probability. He should create a curve like the one shown in Figure 5 during the planning process.

Transmission error rate

The transmission error rate is a function of message size, line conditioning, and hardware characteristics. It is more critical in a centralized data communications environment than in that of a distributed computer network. That is because the former usually does not have error detection or correction features, while the latter does.

As an example, suppose the error rate is not allowed to be more than one in every thousand characters. The designer or planner should calculate the error rate for his candidate system or network. If it is less than 0.001, he is safe. If not, he has two choices. One is to add error detection and correction features to the terminals or terminal controllers. The other is to redesign the system, i.e., use lower speed modems, and fewer terminals per multidrop line, etc.

Traffic bottleneck

After a system has been designed to satisfy specified traffic requirements, the traffic bottleneck of the system initially is not a characteristic of interest. However, if in the future the network has to be upgraded to handle more traffic, knowing where the traffic bottleneck is in the network helps to estimate the incremental cost of expanding the network's traffic handling capacity. For example, if the bottleneck is the high speed line between a concentrator and the central computer, it is quite simple to upgrade the network capacity, either by adding an additional line or by using modems of higher speed. On the other hand, if the bottleneck is at the central computer, the upgrading would be costly.

SELECTION OF COMMUNICATION DEVICES

With the advancement of solid state electronics, communication devices become ever more versatile and generous with options. There are numerous possible combinations available for improving performance, reducing communications cost and satisfying special requirements. However, these goals are not easily achievable. One must know what a vendor has "not" said, what devices are most effective for specific network structures or performance requirement, and how many of each particular device should be used.

A high speed modem can increase throughput, improve response time, and may even reduce costs, but one has to be aware that the transmission error rate is much higher on a line with a higher speed modem. Sometimes, this high error rate prohibits the use of high speed modems with non-intelligent terminals. A modern sharing unit can save cost in a system in which several terminals may be co-located in a same office and share the same multidrop line, but then network reliability will be lowered. On the other hand, a combination of modem sharing units and port or line sharing units may improve reliability as well as reduce costs. A multiplexer or a concentrator may reduce overall communications cost, but only under particular traffic and operational environments. Without proper planning, the introduction of multiplexers and/or concentrators will degrade network performance, and even increase the costs. For more details, refer to Reference 2.

SELECTION OF TRANSMISSION FACILITIES

All data communications network need transmission lines. The planner must choose between dial-up and dedicated lines, choose the right line speed, and calculate and compare line costs.

In the dial-up case, a terminal needs to be connected to the network by dial-up only when there is a need for actual communication. In the dedicated line case, a terminal is connected to the network via a leased or private line and is connected regardless of whether or not there is any actual transmission taking place at a given time. Dial-up connections are used if terminal locations are not fixed (like traveling salesmen's portable terminals), terminals do not belong to a same organization (like in a time sharing environment), terminals are used in remote batch environment with low utilization, or terminals are sparsely located with low utilization. Presently, the dial-up arrangement is either by direct distance dialing (DDD), foreign exchange, or WATS lines. (For the definitions and applications of these three, see References 5 and 6.)

The choice of terminal speed depends mainly on the application, performance requirements, available line tariffs, and cost. It may be determined by rules of thumb, experience, or standard speeds of terminals for the specific application. The simulation program described later and in Reference 4, can be used to determine response time given specific line speeds to see whether a specific speed satisfies a given requirement, and to evaluate cost/performance tradeoffs using different line speeds. In the same network, different terminals and communication devices may be connected to lines with different speeds.

Before the proliferation of specialized common carriers, such as MCI, it was relatively easy to select and interface a specific line type and to calculate line costs, since the user had the convenience of interfacing and dicting with a well defined transmission facility and common carrier. Now, with existing and forthcoming special common carriers of various types, numerous new line tariffs and modifications to old tariffs (AT&T's HI/Lo Density, Digital Data Service, etc.), and the addition of domestic satellite communications service.
and digital transmission service, users are confused. (Communications consultants are happy, however.) The determination of line costs and optimal network topologies is harder than before. The least line cost for routing a line connecting two terminals or devices in a network is no longer necessarily a direct connection, or the one of minimum distance. In some situations, hand calculation is impossible and a special computer program is necessary. For more detail, see Reference 3.

NETWORK STRUCTURES

Network costs and performance depend greatly on the structure the planner chooses. In general, there is no easy way to determine a best structure. Repetitive simulation and design are required to determine the most desirable one. The following are the most commonly used network structures:

1. Point to point connection via dial-up.
2. Point to point connection with leased lines.
3. Multipoint tree-structured connection. The tree structure usually terminates at a multiplexer, concentrator, message switching processor or central computer. In general, the tree network and terminals are so connected that when one device is transmitting a message, it is transmitted to every other device connected in the tree. (This is in contrast with the ring structured network.) However, only the ones with proper identity would actually receive the message.
4. Multipoint ring structured connection. In this structure, terminals or devices form a ring or loop as shown in Figure 2. When terminal A is sending a bit to terminal B, no other terminal knows anything about this bit. (This is in contrast with the tree network.) If this bit information is not for B, B then passes it to C, and so on. At least one of the devices on the ring is a computer which is either the central computer or is a message switching computer capable of switching a message from one ring to another.
5. Multiplexed structure. Several low speed point to point lines or multipoint tree structured lines are terminated at a multiplexer; the multiplexer is then connected to a distant multiplexer with a high speed line. By this arrangement, every low speed line is connected to the distant site as though each had its own line, rather than sharing with other lines.
6. Hierarchical ring structure. Lower level rings are connected to higher level rings via a message switching computer.
7. Hierarchical structure (without rings). Terminals are connected to concentrators or message switching computers first. These computers are then either connected directly to a central computer or connected among themselves.

DESIGN TOOLS

Due to the complexity of many data communications networks, hand analysis and design becomes almost impossible. Computer programs for various analysis and design functions are essential. The usefulness of such programs relies heavily on how convenient it is to repetitively run the programs. Thus, efficiency in size and running time is as important as accuracy. The following is a list of important
design tools:

**Contention simulation program: Determining blocking probability**

This program is useful for dial-up terminals. The output of the program should give the distribution of blocking probabilities for each of the following: number of users (terminals), number of ports, distribution of number of calls per unit time at each terminal, call holding time distribution at each terminal, and distribution of time intervals between two consecutive attempts to obtain a channel for the same call.

**Central processor system configuration program: Verifying performance for specified CPU configurations**

For specified CPU and peripheral device types, the program should indicate whether the specified configuration satisfies throughput and/or response time requirements. If so, the total time span spent by a message in the CPU and peripheral devices will be given. If not, the bottleneck causing the oversaturation will be indicated. Analytic queueing models, as opposed to brute force simulation and analytic closed form formulae, should be used to develop this module. Brute force simulation of the CPU is too complicated and too time consuming, with respect to both development and execution.

**Network simulation program: Simulation of the whole data communications system**

Given a network configuration and traffic requirements, the module should be capable of supplying terminal response time statistics to provide a response time—throughput relationship as a function of network configuration. This is done by simulating a whole system, including regular and intelligent terminals, multidrop lines, concentrators, trunk lines, DCP's and CPU's. For effectiveness and efficiency, simulation, analytic formulae, queueing models, and empirical distributions should be judiciously mixed into the program. Specifically, intelligent terminals, concentrators, and CPU's are to be simulated by queueing models or are to be described by empirical distributions. Lines and terminals are to be simulated.

**Network design program: Concentrator and multiplexer allocation, terminal clustering, multidrop line topological design and economical analysis**

Given the locations of terminals, concentrators, and CPU's; their basic characteristics; the traffic characteristics; and line utilization requirements; multiplexer and concentrator locations are selected; terminals are connected to the proper concentrator or CPU via a multidrop line in a cost effective manner. Thus, an important design goal is to implement optimal heuristic algorithms into the program so that the program size and running time can be proportional to the number of terminals. (In general, the size and running time grow quadratically or cubically with the number of terminals.)

**Network reliability/availability program: Calculation of network reliability**

Given element failure rates (or MTTF's and MTTR's), the module calculates network reliability criteria. A combination of simulation and analytic techniques should be used to ensure the effective determination of reliability for networks with thousands of terminals within reasonable computer time. For more details on design tools, readers are referred to Reference 4.

**DESIGN STRATEGIES AND COST/PERFORMANCE TRADEOFFS**

It is the network planner's responsibility to assist users in defining their requirements and to design a least cost network while satisfying the requirements. To do a good job for a large network, one needs to develop a set of curves to weigh and compare the tradeoffs for cost/performance and for design alternatives. Design tools described in the previous section are extremely useful for this purpose.
Evaluation of design alternatives

By choosing some of the network structures given earlier and by using different line speeds, one can develop several sets of curves like those shown in Figure 3 and Figure 4. For a specified response time requirement, each curve in Figure 3 represents the cost-throughput relation for a specific network structure. For a specified network throughput, each curve in Figure 4 shows the cost-response time relationship for a specific network structure. From these curves, one can determine the most cost-effective network structure.

Evaluation of cost-throughput tradeoffs

Figure 5 shows the relationship between cost and throughput for a fixed response time requirement. With these curves, the network designer can help users to decide how much they are willing to, or must, pay for throughput in the network.

Evaluation of cost-response time tradeoffs (CPU response time, network response time, or CPU and network response time)

In Figure 6, each curve represents the cost-response relationship for each specified throughput requirement. These curves help the user to determine how much he is willing to pay for the response time that he will get, and help the designer to determine the best combination of CPU configuration and line configuration.

Evaluation of cost-reliability tradeoffs

Reliability is another term that users do not quite know how to define. There may be many schemes for improving network reliability. To evaluate them, one must develop a curve to show the incremental cost one must pay for improved network reliability.

Derivation of throughput-response time relationship

Since traffic estimates are rarely accurate and future growth is even harder to predict, network planners need to know how sensitive network performance is to traffic variations from projected levels. Figure 1 shows the relationship between performance and throughput, with performance measured in terms of response time.

REFERENCES
