Management Adviser

Volume 10 | Number 6 Article 5

11-1973

Cost-Performance Tradeoffs in Real-Time Systems Design

Barry E. Cushing

David H. Dial

Real-time EDP systems are the wave of the future, these authors believe. But real-time systems can be incredibly fast or comparatively slow. Speed costs money, and it can cost too much in some instances where a slower system would do the job. So a balancing of needs against capacities is recommended—

COST-PERFORMANCE TRADE OFFS IN REAL-TIME SYSTEMS DESIGN

by Barry E. Cushing
The University of Texas

and David H. Dial
Branch & Orcutt

THE INCREASING use of real-time computer systems in business and other administrative functions presents a new set of opportunities and problems to those managers concerned with getting the most out of expenditures on data processing. Real-time systems are being applied to manufacturing data collection, production scheduling, credit checking, airline and other travel reservations, sales order data entry, bank teller operations, and management simulation. The trend toward more efficient and reliable computer hardware and software, which is often less expensive than that which it replaces, is likely to increase the number and variety of real-time applications. These developments underscore the need for managers to become more familiar with the concepts and technology of real-time systems.

The purpose of this article is to discuss some important aspects of the design of real-time computer systems. The primary objective is to develop an understanding of the trade offs which must be made in the design process between the conflicting objectives of cost minimization and performance maximization. A definition of real-time systems is offered and the essential elements of real-time systems are reviewed. Cost-performance factors in systems design with respect to each of the basic elements are examined in turn. Our goal is to provide managers, system designers, and accountants with a framework for understanding problems of real-time systems design.

Real-time systems

A real-time system may be defined as a data processing system in which the time interval required to process and respond to input data is so small that the response itself is immediately useful in controlling a physical activity or proc-

ess. The most important concept in the definition is that of response time. Real-time systems are sometimes associated with immediate response. However, the length of response time which will qualify a given system as real time is actually dependent upon the nature of the physical activity being controlled by the system. If the activity is the launching of a space satellite, a response time measured in fractions of a second is necessary in order for the system to effectively control the activity. If the activity involves a business function, a response time of several seconds or even a few minutes may be adequate for control purposes. Thus, the nature of the activity being controlled determines the response time which is necessary in order for control to be accomplished by a real-time system.

For a real-time business system, the required response time must be determined for each particular application. For unlike process control systems which direct mechanical devices, the real-time business system controls the actions of human beings. A response time of less than one second is unnecessarily fast. Response times of more than 15 seconds may be too long.

Five elements in system

There are five basic elements of a real-time computer system. These are: (1) on-line direct-access files for storage of system data; (2) one or more central processors; (3) data terminals which provide the interface between the system and its users; (4) a data communications network which links the processor with the terminals; and (5) a software system, consisting of programs, documentation, and other user aids which enable users to operate the system effectively. A diagram of the elements of a real-time system and their relationship to each other is shown in Exhibit 1, page 31. Though not specifically illustrated in the exhibit, the element of software is inherent in each of the other four elements of the system. Each of these elements is discussed in turn in this article.

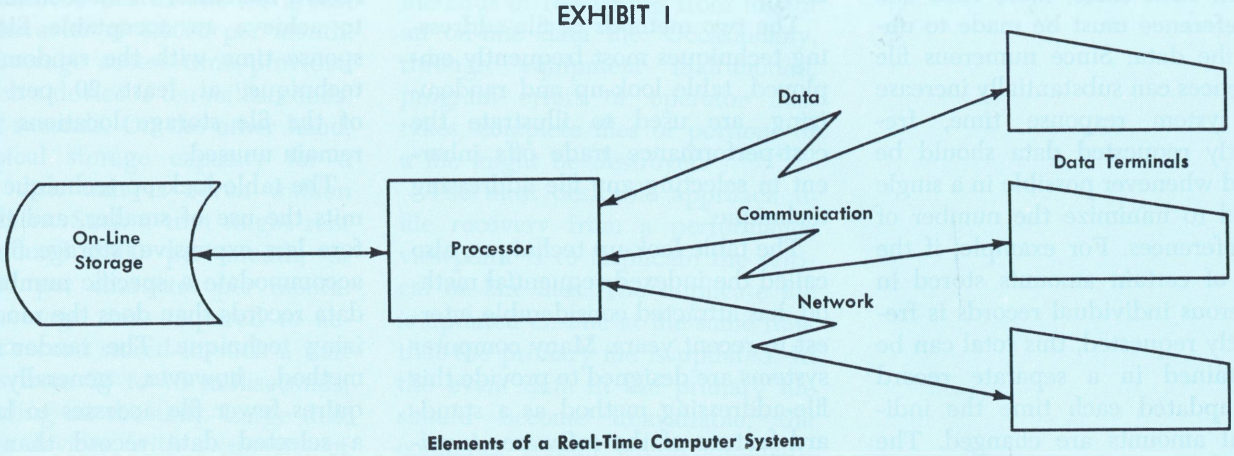
In discussing cost-performance trade offs with respect to real-time systems, it is necessary to clarify the concept of performance. There are two basic performance parameters in a real-time system: response time and reliability. Response time is basically the average elapsed time between data entry and system response. Reliability encompasses both avoidance of system breakdowns and accuracy of data processing. Other performance parameters may be significant in particular applications of real-time systems. Examples include system availability, convenience of working with the system for human operators, and auditability of the system. The systems design process should seek an optimal trade off between cost minimization on the one hand and performance maximization with respect to these objectives on the other.

Application governs time

For a real-time business system, the required response time must be determined for each particular application. Unlike the process-control system which directs mechanical devices with response times measured in fractions of a second, the real-time business system controls the actions of human beings. Response times for interactive accounting and management information systems are measured in seconds or even minutes. When people are operating terminals to interact with a real-time system, the response times must be geared to human reaction times. A response time of less than one second is unnecessarily fast. Response times in excess of 15 seconds, however, may be so long that human operators become impatient. When the operator is engaged in a complex conversation with the computer, the response time needs to be relatively short. Some airline reservation systems, for example, are designed to react to 90 per cent of the transactions in less than three seconds.

Reducing response time will normally cause an increase in the cost of the system since more complex and expensive hardware is required. Increasing the response time may destroy some of the benefits expected from implementation of a real-time system. The importance of the response time

EXHIBIT I



Elements of a Real-Time Computer System

becomes evident when the cost-performance trade offs in various components of the system are examined. A number of factors affect the response time of a real-time system, including the number of operating terminals in the system, the number of messages awaiting processing, the amount of computation required for each message, the speed of the central processing unit, the type of telecommunication network, and the response time of the file storage device.

The degree of reliability required within a real-time system is also dependent upon the particular application. In some applications it is essential that the system be "up" at all times, whereas in others an occasional breakdown may not be critical. In the latter case, however, it may be important to minimize the frequency and/or duration of system breakdowns. With respect to system accuracy, the nature of most real-time applications justifies design of a system which is as reliable as possible. Currently available hardware is highly accurate, and so most of the design problems relating to data accuracy concern the software system.

To achieve a highly reliable system requires duplication of some hardware and procedures and more elaborate hardware and software. These elements may add significantly to the costs of developing and operating a real-time system. However, if the system is less reliable than it should be, expected benefits will not be achieved and actual harm may be done to the organization. Careful analysis of the trade offs affecting system reliability is, therefore, essential in the design of real-time systems.

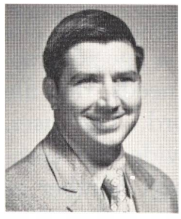
Real-time systems generally require a considerable amount of random access storage capacity. Since most transactions in a real-time system require the computer to access data in the storage files, the response time of the system depends largely on the response time of the storage devices. Certain techniques are available to reduce the response time of a random access file; however, most of these techniques also increase the storage capacity requirements. Such increases in storage capacity add significantly to the cost of the real-time system.

Since most real-time information systems are closely linked to the daily operations of the business, reliability of the storage files is an important factor in system performance. Increased reliability, however, usually means increased system costs for hardware and software. By properly analyzing the cost-performance trade offs in random-access storage files, the system designer can minimize the cost of the data storage and insure adequate system response time and system reliability.

The response time of the file storage device in a real-time system is based upon the number of records which must be accessed by the system before a message can be transmitted to the terminal operator, and upon the time required by the storage device to access a record in the file. Both the number of file references and the average file reference time involve cost-performance trade offs.

Number of file references—The number of file references required to assemble the information to be transmitted as a single message to the terminal operator depends upon two factors. First, the required data may be stored in more than one file

BARRY E. CUSHING, CPA, is associate professor of accounting at the University of Texas, Austin. He has also taught at the University of Illinois and Michigan State University, where he received his Ph.D. Dr. Cushing is the author of numerous articles on management science and information systems as well as a book currently in press, **Accounting Information Systems and Business Organizations,** Addison-Wesley Publishing Co., due early 1974.



DAVID H. DIAL, CPA, is a partner in the firm of Branch & Orcutt, Dallas. Previously he was a management consultant with Peat, Marwick, Mitchell & Co., and director of data processing for the U.S. Navy Data Center. Mr. Dial received his B.B.A. and M.P.A. from the University of Texas, Austin. He is a member of the management services committee of the Dallas Chapter of the Texas Society of CPA's and a member and frequent speaker at technical sessions of the National Association of Accountants.

or in several records within a single file. In such cases, more than one file reference must be made to obtain the data. Since numerous file references can substantially increase the system response time, frequently requested data should be stored whenever possible in a single record to minimize the number of file references. For example, if the total of certain amounts stored in numerous individual records is frequently requested, this total can be maintained in a separate record and updated each time the individual amounts are changed. The cost-performance trade off in this example involves comparing the sum of the cost of storing and additional total figures and the cost of increasing the processing time to update the file containing the total records to the benefit of faster response time.

The second factor affecting the number of file references is the file addressing technique used to locate a specific record in a file containing thousands or even hundreds of thousands of records. File addressing techniques often present distinct examples of cost-performance trade offs in file storage systems, since these techniques influence both the file response times and the file capacity requirements.

Code identifies records

In a real-time information system, each data record is identified by some unique code. For example, the data record for an inventory item might be identified by the inventory part number. Each data record is stored in a separate location in the random-access file. The file locations are also identified by unique numbers, often called file addresses. To access a specific record, the computer needs the file address of the record. The user, however, might provide the computer an inventory part number or a bank account number rather than the file address of the record. The purpose of a file-addressing technique is to provide the computer a method of locating a specific record using only the identifying information supplied by the terminal operator.

The two methods of file-addressing techniques most frequently employed, table look-up and randomizing, are used to illustrate the cost-performance trade offs inherent in selecting any file addressing technique.

The table look-up technique, also called the indexed-sequential method, has attracted considerable interest in recent years. Many computer systems are designed to provide this file-addressing method as a standard feature of the software. A primary characteristic of this technique is the use of one or more tables to provide an index to the random access file. At least two file references are required to locate the desired record: one file reference to read the appropriate indexing table and a second file reference to read the actual data record. With very large data files, the table look-up technique may require a hierarchy of indexing tables. In this case, several file references will be required to read the appropriate table at each level and finally to read the desired data record. Another factor influencing the file response time with the table look-up technique is the time required by the central processing unit to search the indexing tables for the desired entry after the tables have been read into memory from the random access storage file.

Another method of file addressing that is frequently used is a technique called randomizing. The randomizing method transforms a reference number into a random number within the range of file addresses where the desired record is located. This random number is the first address accessed to find the selected record. If the record is not located at the randomized address, another attempt must be made to locate the record at an overflow location. In some instances, several file locations must be accessed to locate the desired record. An important characteristic of the randomizing technique is that as the file packing density increases, the average number of file references required to locate a specific data record increases. In most instances, to achieve an acceptable file response time with the randomizing technique, at least 20 per cent of the file storage locations must remain unused.

The table look-up technique permits the use of smaller, and therefore less expensive, storage files to accommodate a specific number of data records than does the randomizing technique. The randomizing method, however, generally requires fewer file accesses to locate a selected data record than the table look-up method and thus permits a faster response time. When the randomizing technique is used, the average number of file references required to locate a data record can be reduced at the cost of providing a greater percentage of unused storage locations in the file.

Compromise always necessary

The interrelation among file-addressing techniques, file sizes, and file reference times is an important aspect of random access file design. Every file design requires a compromise between response times and data storage costs. Careful analysis by the system designer of these cost-performance trade offs is required to achieve an optimal balancing of conflicting objectives.

Average file reference time—Regardless of the number of file references required to assemble the requested data, the system's performance can be improved by reducing the average file reference time. Reducing the average file reference time, however, requires a trade off in the cost of the random access files and perhaps in the storage capacity of the file device.

Three types of random access storage devices are magnetic drum, magnetic disk, and magnetic strip. A comparison of the access times, storage capacities, and monthly rental costs of these devices illustrates certain cost-performance trade offs inherent in file storage. An average-sized magnetic drum device provides a four-million-char-

acter storage capacity and rents for about $2,000 per month for a cost per character of $.0005 per month. The average access time provided by such a device is ten milliseconds, or .01 seconds. On the other hand, a typical storage capacity for a small disk unit is seven million characters. Such a unit might rent for around $500 per month, or $.00007 per character per month. The average time required to access a record stored in such a unit ranges from 30 to 75 milliseconds, depending on the unit. Large disk storage devices, with a capacity of 100 million characters, provide similar cost and access time characteristics.

The cost-performance trade offs in selecting file storage devices are further illustrated by the magnetic strip device, frequently called a data cell. The data cell is even slower, but is also less expensive, than the magnetic disk. A typical data cell unit has a capacity of 300 million characters and rents for about $2,500 per month, for a cost per character of $.000008 per month, which is about one-ninth the cost per character of disk storage. However, the average access time for a record in a data cell is 500 milliseconds, or seven to sixteen times slower than disk. An access time of 500 milliseconds may be too slow to provide an acceptable response time for a system which has a high volume of file inquiries and updates.

As these examples illustrate, trade offs exist among the three important characteristics of a file storage device: storage cost, access time, and storage capacity. A satisfactory compromise can be achieved in balancing these cost-performance trade offs only by carefully analyzing the requirements of the system and selecting the file storage device which can provide the required performance at the lowest cost.

File recovery

A final example of cost-performance trade offs in file storage is provided by a comparison of the cost and desirability of various methods of recovering from loss of an on-line data file. Occasionally, through equipment malfunction, program errors or operator mistakes, complete files or portions of a file might be destroyed.

The most desirable approach to file recovery from a performance viewpoint is to duplicate the critical on-line files. The duplicate file is updated on-line at the same time that the primary file is updated. In the event data in the primary file should become unavailable, the computer system would automatically channel further file references to the duplicate file and notify the operator of the malfunction. Since this system requires a duplication of a substantial amount of the hardware and the use of specially developed software, the cost of providing file recovery in this manner is quite significant.

A less expensive technique for file recovery is to prepare a copy of the critical on-line files one or more times each day and to maintain a file of all changes that occur to the on-line files throughout the day. If an on-line file is damaged, the on-line system can be temporarily interrupted while one of the backup files prepared earlier in the day is updated for the transactions that have occurred since the backup file was copied. Since the on-line system is unavailable for a short period, procedures must be available for the system users to follow until the on-line system is again operative. In addition, some method must be available to permit updating the computer files for transactions that occur while the system is inoperative.

The cost-performance trade offs for file recovery require balancing the desired level of on-line service with the cost of providing this service. On one extreme, on-line service might not be interrupted more than a few seconds when a storage file is damaged or when a file device becomes inoperative. On the other extreme, the on-line system might be inoperative for several days or even weeks when a storage file is lost. The cost of a system that pro-

Trade offs exist among the three important characteristics of a file storage device: storage cost, access time, and storage capacity.

vides uninterrupted service is necessarily higher than the cost of a system that provides degrading service following a file breakdown. Thus, another decision involving cost-performance trade offs in file storage must be made during systems development.

Central processor

Selection of the central processor configuration in a real-time system involves a number of complex cost-performance trade offs.

Size of primary storage—One of the most critical factors in real-time systems design is the size of the primary storage, or storage area within the central processor. Primary storage, consisting of either cores or semiconductors, is very expensive, ranging around five- to seven-tenths of a cent per character per month. This is ten to 14 times the cost of drum storage, and 70 to 100 times that of disk. However, if primary storage is too small, system response time may be adversely affected.

Most real-time systems use multiprograming, which means that the system can process more than one program simultaneously, though at any one instant system control is devoted to only one program. Multiprograming increases system throughput, and therefore the greater the degree of multiprograming in a real-time system, the smaller will be the average system response time. However, the degree of multiprograming in a system is often limited by the availability of primary storage. The greater the available primary storage area, the greater is the degree of multiprograming possible.

A second illustration of this relationship involves the concept of "virtual storage." In a multiprogramed system, as one program is being executed the system must provide storage area for all other programs and data which are in process and waiting their turn for the computer's attention. The use of primary storage for this purpose may be very expensive. A way of economizing on storage for this "work-in-process" is to store a portion of it on a high speed disk, drum, or other external storage unit. Programs or program sections may thus be relocated, or "swapped," back and forth between primary and external storage several times during their execution. Systems having this capability may appear to have virtually unlimited storage capacity, and are therefore referred to as "virtual storage" systems.

Though virtual storage systems provide a useful means of economizing on storage costs, these devices have an adverse effect upon response time in a real-time system. This is because the extra time required to swap programs back and forth between external and primary storage increases the average time required to process each user's transaction.

Careful analysis of cost-performance tradeoffs involving primary memory size is required in order to obtain a system having an adequate response time and yet avoid excessive expenditures for primary memory.

Processor configuration—A critical factor in real-time system reliability is the processor configuration. A configuration which consists simply of one central processor will at times cause the system to be shut down due to a failure of the processor. Very occasionally, an error in processing may be made as a result of an error by the central processor. The reliability of a real-time system may be considerably improved by configurations which include more than one central processor.

One example of a configuration which increases reliability in a real-time system is the duplex configuration. This system includes two central processors, with one serving as backup for the other. If a failure occurs in the on-line processor, all work is switched over to the backup processor. In such systems the backup processor is generally used for non-real-time functions at those times when both systems are operational. In addition, in the event of a file breakdown, the backup processor may be used to speed file recovery while the on-line processor continues to handle file inquiries and updates as best it can. The duplex configuration is quite common in real-time systems. It greatly increases system reliability in that the probability of failure of both processors concurrently is much smaller than the probability of failure of one processor.

While this example by no means exhausts the number of processor configurations which may be devised to improve reliability in a real-time system, it does illustrate the trade offs involved. Increases in reliability are achieved by duplication of processors, which may significantly increase the cost of the system. The cost of the additional software required is also a relevant factor. However, these additional costs are partially offset by the additional work that may be performed by backup processors while the real-time processor is functioning properly.

Data terminals

The data terminals in a real-time system are the interface between the system and its users. Therefore, decisions involving the terminal subsystem are often a critical factor in the success of a real-time system. Convenience may be a more essential performance factor than either reliability or response time.

A wide variety of terminal devices is currently available for use in real-time systems. The two major categories are (1) teleprinters or teletypewriters, and (2) cathode ray tube (CRT) or display devices. A comparison of some of the major features of these types of devices illustrates some of the cost-performance trade offs involved in the selection of data terminals.

Teleprinters are generally less expensive than display terminals. A purchase price of from $600 to $3,000 is typical for a teleprinter, whereas display terminal prices range from $1,000 to $10,000. Another advantage of the teleprinter is that it automatically produces a paper copy of all terminal

Data terminals in a real-time system are the interface between system and users.

activity, which in some cases significantly improves the convenience and auditability of the system.

The more expensive display terminal, however, has several performance advantages over the teleprinter. One important advantage is output speed. Typical printing speeds of teleprinters range from 10 to 30 characters per second. In contrast, display terminal output speeds depend upon the transmission speed of the data communication facility, and therefore speeds ranging from 60 to 240 characters per second or more are common. This factor is particularly important if output volumes are large.

Other advantages of display terminals over teleprinters include: (1) easier correction of errors in previously entered data by modifying only erroneous characters rather than retyping entire lines, (2) superior capability in displaying graphic output, and (3) noise-free operation. In addition, some display devices can store in memory more lines of data than can fit on the screen at any one time, in order that the operator can refer back to such data after it leaves the screen. Many display terminals can be equipped with a device which will produce a paper copy of whatever is on the screen when desired. However, all of these additional performance factors add to the expense of the terminal device.

Another critical decision relating to data terminals in a real-time system is the appropriate number of terminals in the system. User convenience is maximized if there is one terminal available for each user. However, this also requires a maximum expenditure on terminals. If several users can share each terminal, the expenditure on terminals may be reduced. However, such a reduction in cost is accompanied by a reduction in user convenience. This trade off involves evaluating the needs of the users relative to the cost of the terminals.

Still another factor relating to the selection of data terminals involves the possibility of using terminals which have a "stand-alone" capability. Such terminals can continue to perform such functions on their own even if the central computer system goes down. For example, some terminals can record and store transaction data on a machine-readable medium for transmission after the failure has been corrected and the system is available. Such terminals may also be capable of preparing a printed record of such transactions if one is desired. To obtain a stand-alone capability may require a more expensive terminal.

Several of these elements in selection of data terminals in a real-time system are illustrated by the case of a hospital which developed such a system for processing patient charges, laboratory test results, and related patient data. Terminals in each laboratory, in the pharmacy, and in other locations from which patient charges originated were used to enter transaction data into the system. Terminals were also located at nurses' stations throughout the hospital so that laboratory test results could be sent to them for inclusion in patient records, and so that doctors could use the terminals for fast retrieval of patient medical data. Still another terminal was located at the accounts receivable office for use in recording patient checkouts and preparing receipts for collections from patients.

The choice of terminals for the nurses' stations presented an interesting situation. Cost minimization was an important objective, and documentaton of laboratory test results was essential. These criteria pointed to the selection of an inexpensive teleprinter, such as the Teletype Model 33 at a purchase cost of $600. However, due to the proximity of the nurses' stations to the hospital rooms, another essential objective was noise-free operation. Furthermore, due to the intended use of these terminals by doctors to retrieve patient data, output speed was very important. For these reasons, a small CRT display terminal with an attached hard copy unit was chosen at a purchase price of approximately $3,500. Though costing almost $3,000 more per unit, this device met all performance criteria, including minimization of machine noise.

The selection of terminals for the pharmacy and for the accounts receivable department also required a compromise of the cost minimization objective. Because these departments dealt directly with patients and the general public, it was considered essential to utilize a stand-alone terminal which could record transaction data and provide receipts even while the central computer system was down.

Data communications

The terminals used to communicate with the computer in a real-time system are often located at some distance from the computer. A telecommunication network is required to link the various terminals with the central computer. Basically, this telecommunication network consists of a transmission link and a set of electronic devices used to increase the efficiency of the network. A well planned network utilizes the combination of transmission links and peripheral devices that provides the required transmission rate and system response times at the lowest cost.

The cost of a telecommunication network is determined by several factors including the line transmission speed and the choice of leased

TABLE I

Computation of Line Cost for a Leased Line

Rate Structure		Detroit-Chicago Hookup		
Miles	Rate/Mile	Computation		Result
1-25	$3.00	25 x $3.00	=	$ 75.00
26-100	2.10	75 x $2.10	=	157.50
101-250	1.50	138 x $1.50	=	207.00
251-500	1.05	.....		.....
501-up	0.75	.....		.....
		Totals 238 miles		$439.50

TABLE II

Computation of Switched Line Cost

No. of Calls	Orders/Call	No. of Orders	Length of Call	Cost/Call	Total Cost
600	1	600	2 minutes	$0.90	$540.00
100	2	200	4 minutes	1.17	117.00
20	3	60	6 minutes	1.71	34.20
10	4	40	8 minutes	2.25	22.50
	Totals	900			$713.70

or switched lines. Certain alternatives to the use of private lines are available such as Private Exchange (PBX) or multidrop lines. Each of these factors affects not only the cost of the network but also the performance of the system.

Line transmission speeds—Communication lines can be classified into three primary categories based upon the number of data bits per second that can be sent over the line. To measure transmission speeds in characters per second, the number of bits required to represent a character must be known. In the following discussion, a ratio of ten bits per character is assumed since this figure is representative of the transmission codes commonly used.

The lowest speed lines, called subvoice-grade lines, are designed for telegraph and similar machines transmitting at speeds generally not exceeding 300 bits per second. A subvoice-grade line can provide a low-cost communication link for a real-time system that uses only typewriter-speed terminals operating at transmission rates up to 30 characters per second.

Voice-grade lines, originally designed for telephone communications, provide transmission speeds as high as 9,600 bits per second. When the regular dial-up telephone lines are used however, the maximum attainable trasmission rate is limited to 4,800 bits per second. The high speeds are possible on private lines that are specially conditioned for data transmission. Real-time systems using display terminals will usually require voice-grade channels to take advantage of the extremely high transmission speed possible between a computer and a display terminal.

Wideband lines provide the capability of transmitting data at speeds up to 500,000 bits per second. One application for wideband lines is high-speed communications between two computers. Subvoice-grade and voice-grade lines are currently the most important communication links for real-time business systems. The speed at which input data can be entered, or output data interpreted, by human operators using keyboard terminals is so severely limited that very high-speed transmission facilities are not usually required.

An obvious cost-performance trade off exists between subvoice-grade and voice-grade lines. Although the voice-grade line costs more to lease than the subvoice-grade line, the voice-grade line permits a substantially greater transmission speed. These performance factors and the cost differentials among various line transmission speeds must be carefully examined to determine the proper balance between line cost and transmission speed.

Leased versus switched lines—Two basic options are available with respect to usage of data communication facilities. These are (1) leased or private lines, and (2) switched lines or dial-up service. Leased lines are devoted exclusively to the use of a single customer. Dial-up involves simply using the long distance telephone service available to the general public. The cost of a leased line between two points is fixed and is determined by the length of the line. The cost of dial-up service is variable with distance and usage time. Therefore, dial-up service is less costly than a leased line—up to a breakeven volume, beyond which the leased line is more economical.

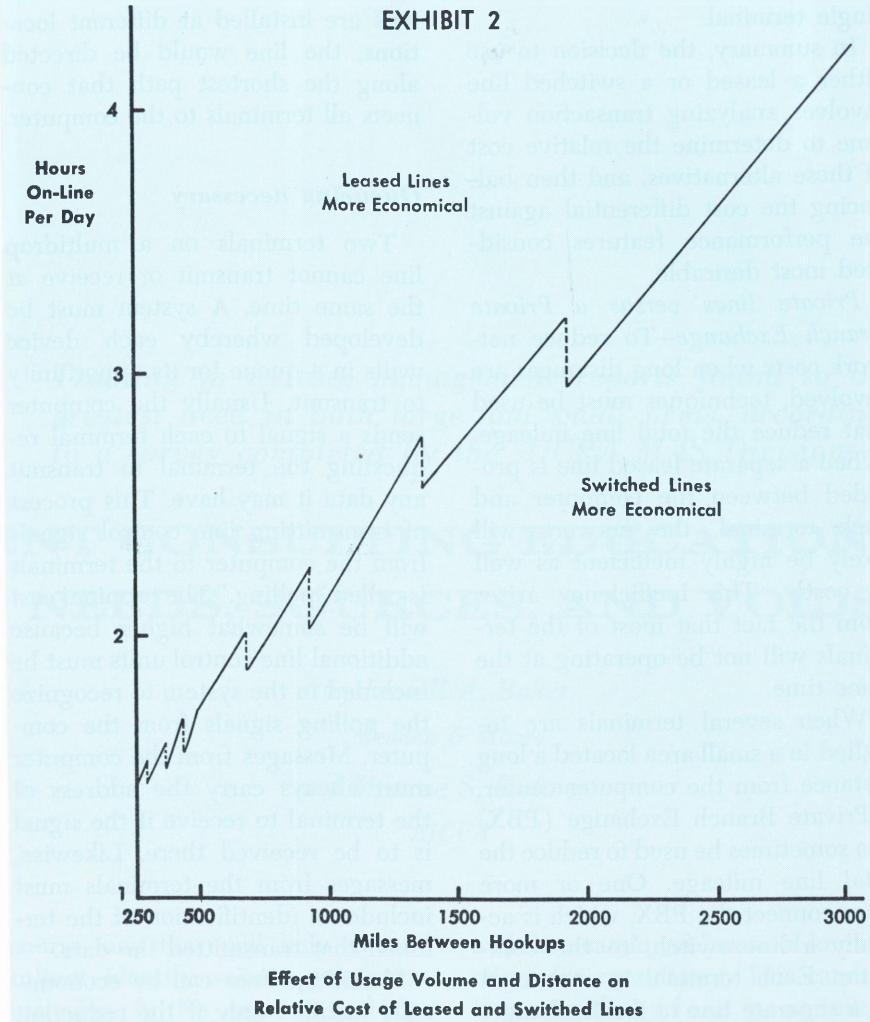
To illustrate the relative cost differential between leased and switched lines, consider the case of a company having a main office and computer center in Chicago and a branch sales office in Detroit. The company wishes to connect a data terminal in its Detroit office to the computer center in Chicago using data communication services, so that sales orders from its Detroit customers may be processed on a real-time basis. If a leased line is used, the differential cost per month will include approximately $30 for communications hardware at the Detroit and Chicago locations, plus the cost of the line itself. A rate structure embodying a decreasing cost per mile as line mileage increases is used to compute the line cost. This computation, using actual Bell system rates in effect June 1, 1973 for a voice-grade line, and assuming a distance of 238 miles between the two locations, is illustrated in Table I, on this page.

The resulting line cost of $439.50, plus the additional hardware cost of $30, yield a total cost per month for the leased line of $469.50.

In contrast, the cost of switched lines (dial-up service) depends upon long distance rates, number of calls, and length of each call. For example, assume that the long distance rate from Detroit to Chicago is $0.90 for the first three minutes and $0.27 for each additional minute. If an average of 900 sales orders per month are received from Detroit, and if each requires a separate call of less than three minutes in duration, the cost will be $0.90 $\times$ 900 = $810 per month. However, if some calls are entered in groups of two or more, the average cost per order will be less due to the lower number of calls required and the smaller rate per minute once a call exceeds three minutes. To compute the actual cost would require knowledge of the pattern of receipt orders at the sales office and the average length of time required to enter order data over the terminal. For example, assuming an average of two minutes connect time per order, Table II, page 36 illustrates this computation under an assumed pattern of receipt of orders.

Exhibit 2, this page, illustrates the relationship of usage volume and distance between hookup points to the breakeven point between leased lines and switched lines. Each point on the breakeven line represents a point where the monthly cost of a leased line for the number of miles given on the horizontal axis is exactly equal to the monthly cost of a switched line for that number of miles which is used for the amount of time per day shown on the vertical axis. The computations underlying the chart incorporate line costs only, and assume a month of 22 working days. Furthermore, the computations are based on the dial-up rate for each additional minute beyond the initial three minutes, which means that the chart reflects a situation in which transactions are entered in large batches (remote batch processing) such that the extra rate for the first three minutes increases total cost by an insignificant amount. If, alternatively,

EXHIBIT 2



Effect of Usage Volume and Distance on Relative Cost of Leased and Switched Lines

transactions are entered as they are received, the pattern of receipt must be known or assumed before a chart of this type may be prepared. Note that the breakeven line itself for any such chart would be almost identical in appearance to the line in Exhibit 2.

To further explain the breakeven line, note that each discrete drop in the line represents a mileage level at which a rate break occurs in the station-to-station dial-up rate. For example, at 676 miles the rate per minute increases from $0.32 to $0.35, causing a discrete drop in the economic desirability of switched lines. Furthermore, the change in the slope of the line at 500 miles reflects the decrease in the cost per mile of a leased line from $1.05 to $0.75 (see Table I). In conclusion, the exhibit demonstrates that for short distances leased lines are more economical unless the volume of usage is quite small, whereas for long distances switched lines are more economical unless the volume of usage is quite high. In any given situation the cost differential between these two alternatives may be quite significant.

In addition to relative costs, the choice between leased and switched lines is affected by such performance factors as transmission speed, error rate, and flexibility. Transmission speed favors leased lines, since 4,800 bits per second is the maximum attainable transmission speed with dial-up lines. Error rates also favor leased lines, which can be conditioned to reduce error rates significantly below those experienced on dial-up lines. However, flexibility favors the use of dial-up service in the sense that more than

one system may be accessed from a single terminal.

In summary, the decision to use either a leased or a switched line involves analyzing transaction volume to determine the relative cost of these alternatives, and then balancing the cost differential against the performance features considered most desirable.

Private lines versus a Private Branch Exchange—To reduce network costs when long distances are involved, techniques must be used that reduce the total line mileage. When a separate leased line is provided between the computer and each terminal, the network will likely be highly inefficient as well as costly. This inefficiency arises from the fact that most of the terminals will not be operating at the same time.

When several terminals are installed in a small area located a long distance from the computer center, a Private Branch Exchange (PBX) can sometimes be used to reduce the total line mileage. One or more lines connect the PBX, which is actually a line switch, to the computer. Each terminal is connected by a separate line to the Exchange. A terminal is connected to the computer only while the terminal is being used. The economic feasibility of using the PBX depends upon whether or not the reduction in line mileage provides sufficient cost savings to offset the cost of the PBX.

One disadvantage of the PBX approach is that terminal operators may at times be unable to obtain a line to the computer because all lines are busy. The number of terminals that can be used simultaneously cannot exceed the number of lines from the Exchange to the computer. Thus, a cost-performance trade off arises as the reduction in network cost must be balanced against the possible reduction in system availability as an operator awaits a line to the computer.

Multidrop versus private lines—Another technique for reducing the total line mileage is to use a multidrop line, a single line connected to several terminals. If the terminals are installed at different locations, the line would be directed along the shortest path that connects all terminals to the computer.

Queueing necessary

Two terminals on a multidrop line cannot transmit or receive at the same time. A system must be developed whereby each device waits in a queue for its opportunity to transmit. Usually the computer sends a signal to each terminal requesting the terminal to transmit any data it may have. This process of transmitting line control signals from the computer to the terminals is called "polling." The terminal cost will be somewhat higher because additional line control units must be included in the system to recognize the polling signals from the computer. Messages from the computer must always carry the address of the terminal to receive if the signal is to be received there. Likewise, messages from the terminals must include an identification of the terminal that transmitted the data.

Multidrop lines can be economically justified only if the reduction in line mileage provides a cost savings sufficient to offset the cost of the additional hardware required in the system. However, any cost saving achieved may be offset by a performance reduction in the form of increased system response time and decreased system reliability. While a message is being transmitted to or from one terminal, all other terminals on the line must wait, which means that system response time is increased in some cases. On the other hand, if one section of a multidrop line fails, the system is unavailable to all users located down the line from that point. A line failure in a system using point-to-point lines or a PBX will generally only affect one user. Thus the cost-performance trade off to be considered with multidrop lines is the reduction in total network costs achieved by increasing average system response time and sacrificing some degree of system reliability.

Many software decisions are inherent in decisions relating to the four areas of hardware already discussed. Examples include the selection of file reference method, the design of file structures, and the selection of processor configurations. These are not discussed further here.

Perhaps the most important cost-performance trade offs involving software in a real-time system relate to the reliability of the system. Software costs are the "personnel costs" of system analysis and programing. Software reliability is dependent upon such factors as the extent of system testing, the adequacy of system documentation, and the thoroughness of input data validation. These factors have been discussed extensively elsewhere in the literature* and are not belabored here.

Summary and conclusions

Cost-performance trade offs are inherent in decisions relating to file storage, central processor, data terminals, data communications, and software in a real-time system. Though for convenience these five topics have been discussed separately in this article, they are closely interrelated in the design process. The decisions made have important implications for such performance factors as system response time, reliability and user convenience.

Real-time systems are the wave of the future in computerized data processing. Therefore it is important that the managers, accountants, and other non-specialists involved in the planning and evaluation of real-time systems develop a general understanding of the performance economics of such systems. Though a comprehensive treatment is beyond the scope of an article of this length, we have attempted to discuss some of the more important cost-performance trade offs in real-time systems design.

*For a comprehensive treatment, see James T. Martin, *Programing Real-Time Computer Systems*, Englewood Cliffs, N.J., Prentice-Hall, Inc., 1965.