Interactive user productivity is a measure of effective communication between man and the computer. Explored in this paper is the relationship between computer response time and user performance, and the separation of user cost from system cost. Strategies for effectively managing installations are presented and discussed.

Interactive user productivity

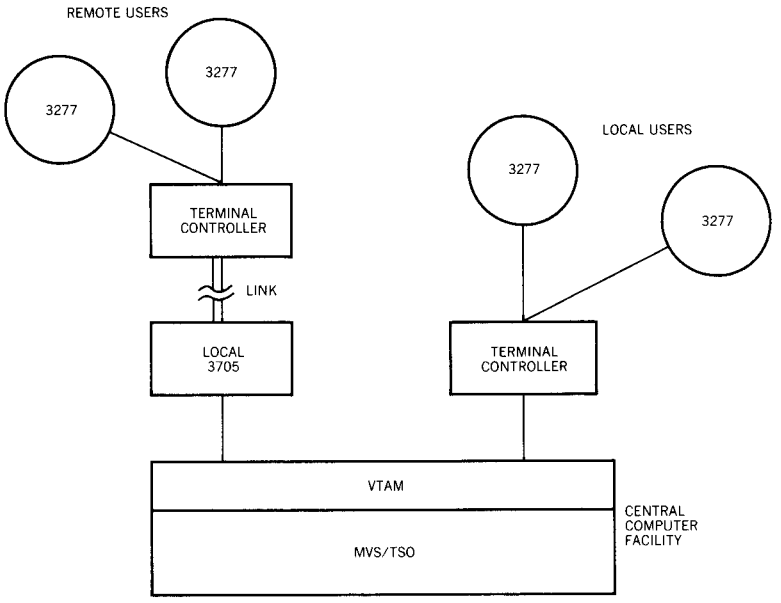
by A. J. Thadhani

The use of computers where users continuously interact with the system began in the mid-1960s. Interactive applications became widespread commercially in the 1970s with the availability of several interactive systems, such as the Time Sharing Option (TSO) on the Multiple Virtual Storage operating system (MVS) and the Conversational Monitor System (CMS) on the Virtual Machine Facility (VM/370).

Although system response time and its effect on user behavior have been widely discussed during the last decade, key issues are still being debated today. Some of these are the acceptable range of system response times from a user's viewpoint, the variation in user performance within that acceptable range, and the shortest response time below which the system no longer limits user performance. Delivering very short response time to interactive users is now technologically feasible. The main problem is the lack of quantification of the value of response time in a suitable form to aid management in making the appropriate tradeoffs between systems cost and user performance.

User behavior has been previously explored in the research environment on VM/370 systems[1] and the IBM Time Sharing System (TSS).[2] However, no work has been reported on user behavior in production MVS/TSO environments and for interactive data base applications. In this paper, the relationship between

Figure 1 Generalized system configuration used for the study



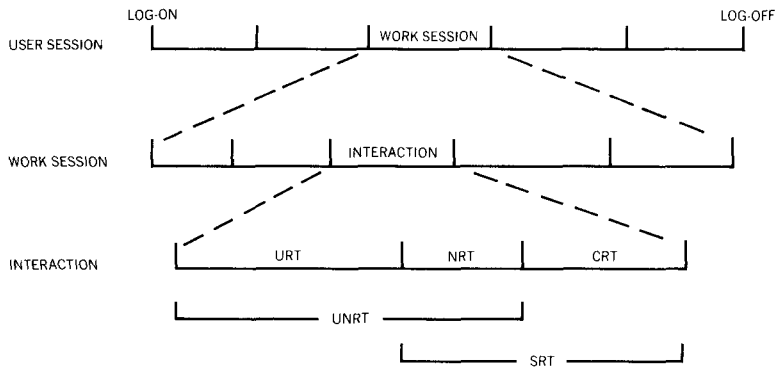
system response time and user performance is examined on two production MVS/TSO systems in IBM. One of these supports manufacturing operations and the other supports software development. Two important aspects of the analysis are the quantification technique and the apparent nonlinear relationship between subsecond computer response time and user productivity.

System configuration

The interactive system that served as the basis for this analysis was TSO under MVS, running on System/370 Model 168 processors. TSO is a general-purpose interactive system described in detail in many publications and user manuals.[3] Data from our two systems, referred to as system A and system B, are used in this analysis. System A is a System/370 Model 168 Attached Processor, supporting engineers and programmers involved in manufacturing operations. System B is a System/370 Model 168 multiprocessor system, supporting programmers involved in software development.

The systems support both interactive TSO and batch applications. The weekday first-shift (8 a.m. to 5 p.m.) workload is mostly TSO, with some batch applications running in the background at a lower priority. On second, third, and holiday shifts, the workload is predominantly batch with little TSO activity.

Figure 2 Relationships between user session, work session, and user interaction



The configuration common to both systems is shown in Figure 1. Users interact with the central computer facility from either local or remote terminals. Local terminals attach directly to channels of the central computer, and remote terminals attach via a transmission link (telephone lines in this case) to the IBM 3705 network controller. Approximately fifty percent of system A users interact from remote terminals, whereas fewer than ten percent of system B users are remote. The IBM 3277 video display station is the predominant terminal type on both systems, and it communicates with the central computer via the Virtual Telecommunications Access Method (VTAM).[4]

user session and work session

Figure 2 shows the relationships between the user session, work session, and user interaction. A *user session* is defined as a series of interactions between the user and the computer. Users enter commands and receive system responses at their terminals. User sessions begin with a log-on and end with a log-off. In this paper, a user session is divided into multiple *work sessions*, each consisting of 100 interactions.

interaction

An *interaction*, consisting of a user command and a system response, can be divided into three time sequences, a user response time (URT), a network response time (NRT), and a computer response time (CRT). *User response time* is the time between a user's receiving a system response and his entering the next command. *Network response time* consists of two delay components—the network delay in transmitting the user's command from the terminal to the computer and the delay in response from the computer to the user's terminal. *Computer response time* is the time between the computer's receiving a user's command from VTAM and the computer's sending its completion response to VTAM for transmission to the user. *System response*

time, as seen by users at the terminal, is the sum of NRT and CRT.[1] The *user and network response time* (UNRT) seen by the computer is the sum of URT and NRT.

computer- and human-intensive

User interactions may be divided into two groups. Interactions that experience long computer delays and consume large amounts of computer resources are called *computer-intensive interactions*. An example is the compiling of a large source program, taking tens of seconds to complete. Interactions that experience short computer delays and consume small amounts of computer resources are called *human-intensive interactions*. Most Edit commands, completing in a few seconds, are good examples of human-intensive interactions.

By definition, the classification of interactions as computer-intensive and human-intensive is a function of computer speed. On a very powerful computer, most interactions would be classified as human-intensive, whereas on a very slow computer most interactions would be considered computer-intensive. We estimate that on the systems we studied approximately ninety-five percent of all interactions were of the human-intensive type.

interactive user work

Users accomplish their goals during a session by means of interactions, which are commands supported by the interactive system. For example, a software developer's goal of creating and executing program A may be accomplished by a series of editor commands to perform the following functions: create the source program; save it on a disk; and compile and execute the program. From the user's viewpoint, each command is an interaction necessary to achieving the goal. *Interactive user work*, therefore, is defined in terms of the number of interactions between the user and the computer system.

interactive user productivity and work session time

Interactive user productivity (IUP) is defined as the number of interactions per user during a one-hour period and is expressed as interactions per user per hour. Thus IUP expresses the user's interaction rate; it is not the rate of completing user-defined tasks. This measure cannot be used in comparing the efficiency of two users doing the same task with different numbers of interactions. It is, however, a useful measure of the average interaction rate of users currently using the system, particularly when the aggregate user work pattern is invariant.

Interactive user productivity depends on two factors: user capability and system capability. User capability varies among individuals and is a function of typing speed, thought process speed, concentration level, time of day, etc. System capability is a function of processor speed, system configuration, batch workload, bottlenecks, etc. The man-computer system is said to be balanced when the system capability of processing user interac-

tions matches the aggregate user capability of generating interactions. Otherwise, it is unbalanced. It is either user-limited if the system can process more interactions than are generated or system-limited if the system prevents the users from generating interactions at their capability levels.

User work session time is defined as the time for a user to accomplish one hundred interactions with the computing facility. Interactive user productivity and work session times are inversely related, with higher user productivity associated with lower work session times and vice versa.

Methodology

The data for our analysis were extracted by the Resource Management Facility (RMF), an MVS measurement and analysis tool designed to monitor selected areas of system activity over an installation-specified time interval.[5] A fifteen-minute interval was specified on both systems. System A data were collected during March and April, 1980, and system B data, from July through September, 1980. Neither new applications nor major software changes were installed during the measurement periods.

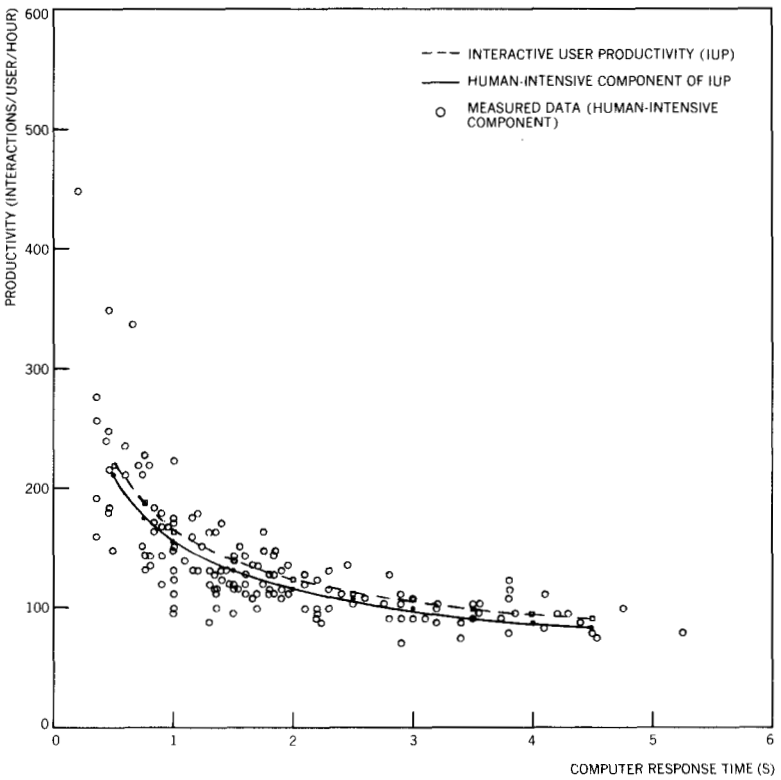
The MVS/TSO systems were installed with three periods for TSO interactions, each period being defined by a *service unit* threshold. (An MVS service unit is a measure of computer resource usage.) All TSO interactions begin in period 1, with high priority for computer resources. Interactions not completing in period 1 and consuming more service units than the period 1 threshold are moved to period 2 and given a lower priority. Similarly, interactions move from period 2 to period 3. Service unit thresholds were set so that approximately ninety-five percent of all TSO interactions (human intensive) complete in period 1.

Among other system statistics, RMF collects the number of completed interactions, resources consumed, and computer response time for each TSO period. The number of logged-on users is also captured. Both systems automatically log-off inactive users, i.e., those who have not interacted during the past thirty minutes. Therefore, the number of logged-on users is a measure of the number of active users.

Interactive User Productivity (IUP) is computed as follows:

$$\text{IUP} = \frac{\dfrac{\text{interactions}}{\text{interval}}}{\dfrac{(15\text{ minutes})}{\text{users}}} \times \frac{\dfrac{\text{intervals}}{(\text{four})}}{\text{hour}} \quad \text{interactions/user/hour.} \qquad (1)$$

Figure 3 Interactive user productivity versus computer response time for human-intensive interactions for system A



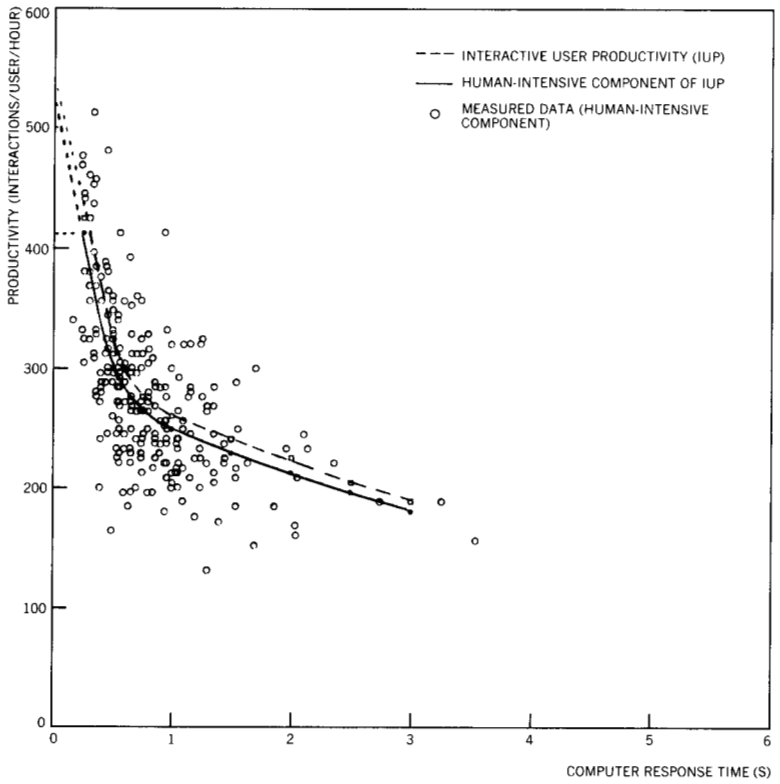
The human-intensive component of IUP is computed by using completed period 1 interactions, instead of all TSO interactions in Equation 1.

When the number of logged-on users on the system is small, it is possible for a few users to have an inordinately large effect on the aggregate user work, and hence bias the results. To minimize bias, all data with fewer than twenty-five logged-on users were excluded from the analysis. Furthermore, to minimize the effect of changes in the aggregate user work at different times of the month, the data collected were separated into groups of six to eight consecutive days and analyzed separately. One representative sample is used for IUP for each system, with least squares fitted third-order negative exponent polynomial for IUP.

Results and their interpretation

The data summarized in Figures 3 and 4 show that interactive user productivity and the computer response time (CRT) for

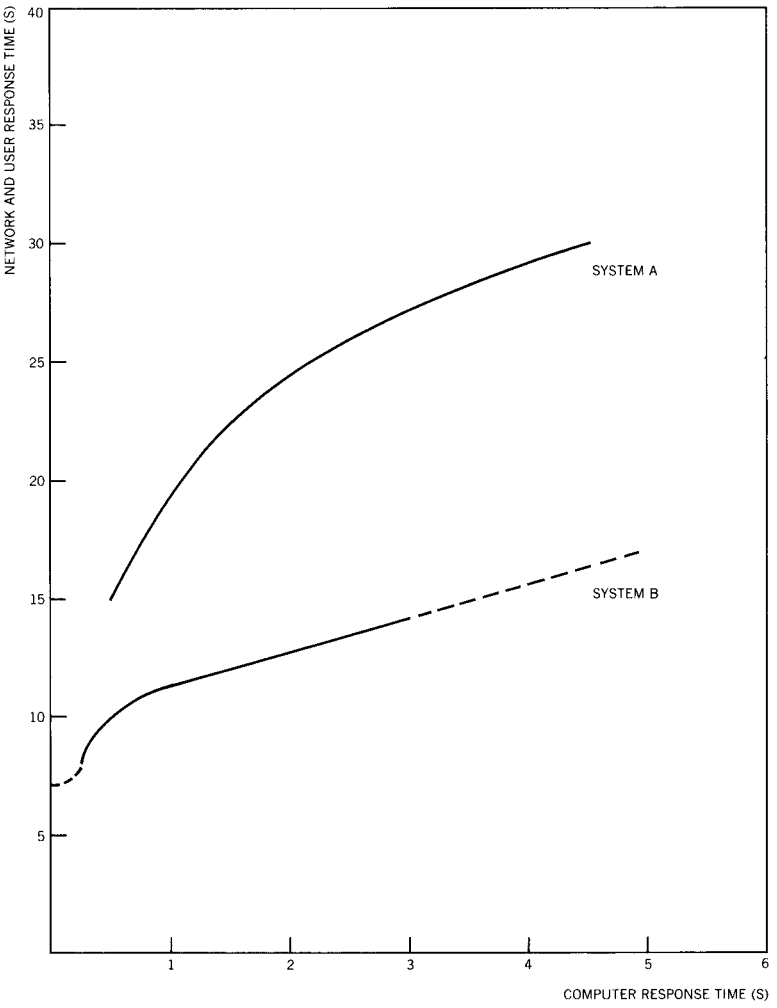
Figure 4 Interactive user productivity versus computer response time for human-intensive interactions for system B



human-intensive interactions are related. The slopes of the curves are significantly larger in the 0.25-second to 1.0-second CRT range than for values of CRT greater than 1.0 second. On system A, the IUP of 222 interactions/user/hour at 0.5 second is over two times larger than the IUP of 106 interactions/user/hour at 3.0 seconds. On system B, the IUP is 67 percent larger at 0.5 second than at 3.0 seconds.

Most of the data for system B are in the CRT range of 0.25 second and 3.0 seconds. Less than two percent of the data are in the zero-second to 0.25-second range. The data, therefore, are not strong enough to conclude that 0.25 second is the limiting CRT at which users reach maximum productivity. We know, however, that human capability is limited. The data do suggest, however, that the limiting CRT lies somewhere between zero second and 0.25 second. Two extrapolations, one tangential and one constant, bound the maximum user capability between 420 and 530 interactions/user/hour, as shown by the dashed extension in Figure 4. Note that a zero-second CRT does not correspond to instantaneous system response time at user terminals. Network

Figure 5 Relationship between user and network response time and computer response time for systems A and B



delays, though minimal for locally attached terminals, can range from 0.02 second to 0.4 second, depending on the terminal configuration and loading factors.

UNRT and CRT

The RMF does not collect statistics on user response times and network delays. Therefore, the sum of user and network response time (UNRT) is computed by equating the time for an interaction (UNRT + CRT) with the reciprocal of interactive user productivity, as shown in Equation 2.

$$\text{UNRT} = \frac{3600 \text{ seconds}}{\text{IUP}} - \text{CRT seconds}. \tag{2}$$

UNRT for both systems are aggregate UNRT for local and remote users. The data as displayed in Figure 5 show the way in which UNRT is related to CRT. The slope of the curve is significantly larger in the 0.25- to 1.0-second CRT range than for values of CRT greater than 1.0 second. Since network and transmission delays for remote users are large and since approximately fifty percent of the users studied were remote, NRT was a significant component of UNRT for system A. By comparison, most system B users were locally attached with minimal NRT. Thus UNRT on system B closely approximates URT and is more representative of actual user performance.

In the one-to-three-second range, the relationship between UNRT and CRT for system B is similar to the relationship between URT and SRT reported for the TSS system.[2] For a one-second increase in CRT, there is a 1.4-second increase in UNRT. In the zero-to-one-second range, however, the relationship is nonlinear. UNRTs were 8.3 seconds and 11.4 seconds at CRTs of 0.25 second and one second, respectively. The two extrapolations of interactive user productivity, discussed previously, bound the minimum UNRT between 6.8 seconds and 8.3 seconds.

interpretation of results

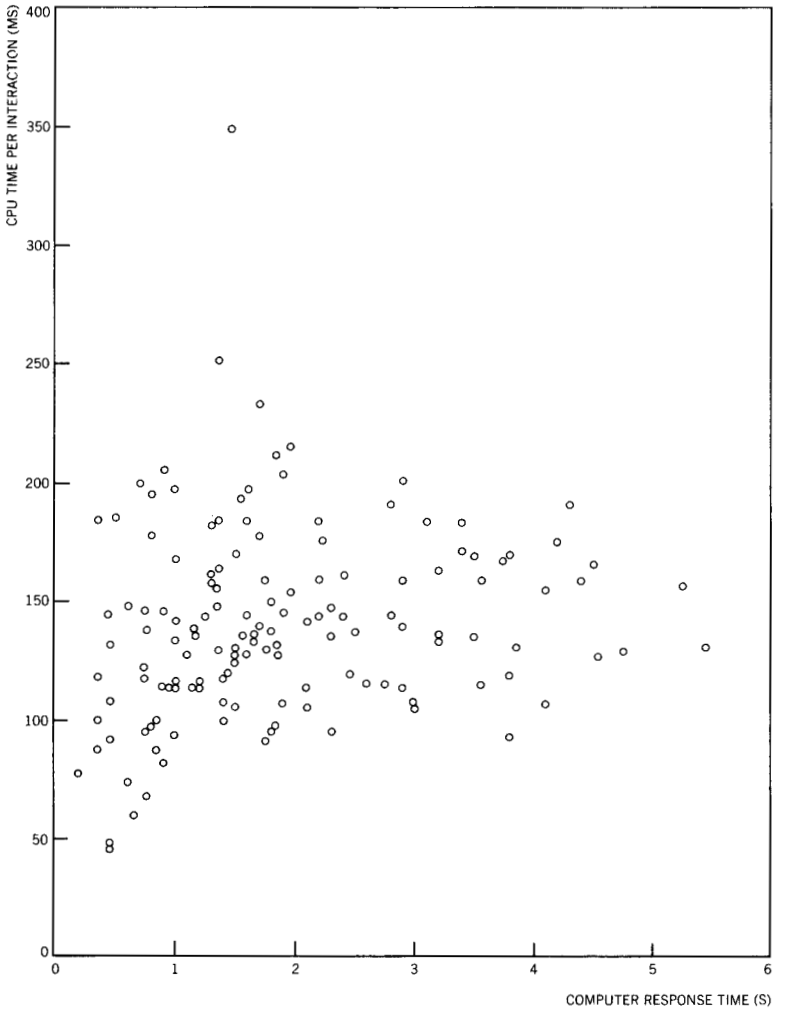
The data show that interactive user productivity is larger at short computer response time. The psychological explanation of this phenomenon is based on the functioning of human short-term memory.[6] The following is Doherty's[1] explanation of the human behaviorist view:

> "The traditional model of a person thinking after each system response appears to be inaccurate. Instead, people seem to have a sequence of actions in mind, contained in a short-term memory buffer. Increases in SRT seem to disrupt their thought processes, and this may result in having to rethink the sequence of actions to be continued."

Based on this concept, the UNRT-CRT relationship indicates that the greatest disruption to human thought process and user capability occurs in the subsecond CRT range.

There could, however, be other causes. The fact that IUP and URT are correlated with CRT does not necessarily imply that higher CRT causes higher URT and lower IUP. Higher URT at higher CRT could occur if users modify their behavior and issue few complex interactions instead of several simple interactions.[7] On both systems, however, there is no significant change in the human-intensive component of IUP. Furthermore, Figure 6 shows that the CPU time per interaction, a computer measure of interaction complexity, shows no pattern of increase in CPU time at larger CRTs. Though not conclusive, the data suggest that no significant change in user work pattern occurred in the 0.25-second to 3.0-second range.

Figure 6 CPU time per interaction versus computer response time for system A



comparison between systems

There is a wide difference in user performance on the two systems. UNRTs on system A are larger than for system B, at all values of CRT. Part of this difference may be due to the different types of work done on the systems. Work done on system A may be more complex; hence, the URT component of UNRT may be larger. The other difference, as discussed earlier in this paper under the heading of UNRT and CRT, is in the NRT component.

The network response time component of UNRT for remote users is generally an order of magnitude larger than for local users. This difference is explained by comparing the time for a simple scroll command. NRT for a scroll consists of the time to establish

communications and the time to transfer a full screen of data. Protocols to establish communication between remote terminals and the computer take approximately one second. For local terminals, protocols are insignificantly in the tens of milliseconds range. A full screen write on the 3277 terminal 80-by-24-character screen takes two seconds for remote terminals over a 7200-baud line. For local terminals at channel speeds, the same action requires approximately 0.15 second. NRT for scroll is, therefore, approximately 0.15 second for local terminals and 3.0 seconds for remote terminals, which is a significant difference.

In summary, the data show that IUP is correlated with CRT, with significantly larger IUP in the subsecond CRT range. Examination of computer parameters suggests that there is no significant change in user work in the 0.25-second to 3-second CRT range. This suggests that humans are more efficient at short CRT. Finally, local attachment of terminals is preferable to remote attachment, since large network and transmission delays prevent subsecond system response time for remote users.

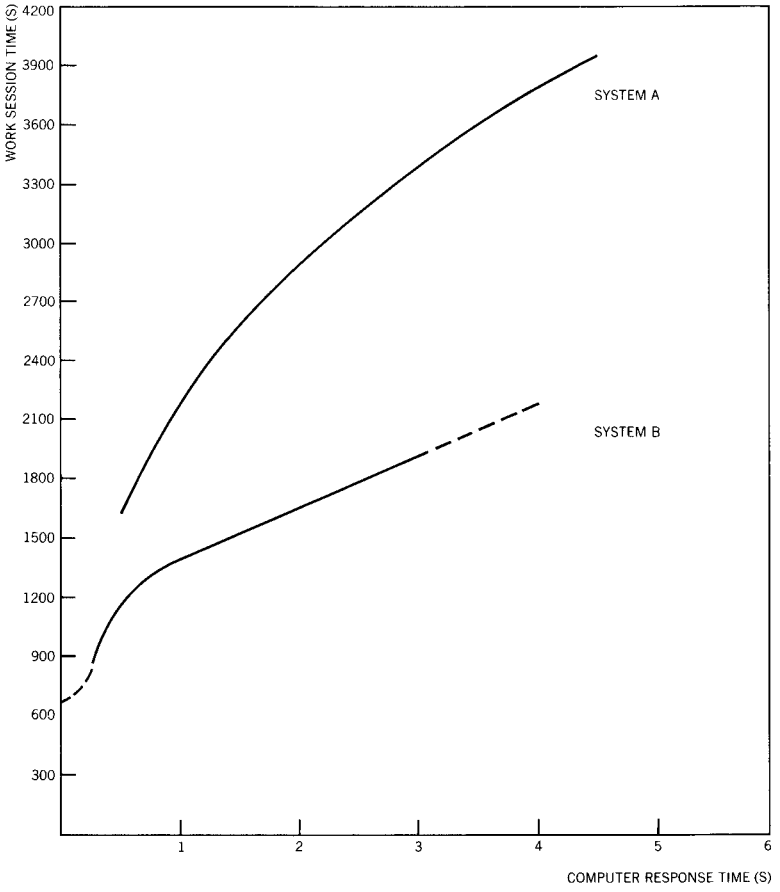
work session time

Figure 2 shows the relationship between user session, work session, and interaction times, where a user session is divided into multiple work sessions. For purposes of this analysis, a *work session* was arbitrarily defined earlier in this paper as consisting of 100 interactions. In deriving work session times, it is assumed that user work, an example of which—expressed in human terms to be the correct execution of program A—requires a series of interactions that are independent of the CRT. *Work session* time for 100 interactions is computed as follows: 100 × (UNRT + CRT). The UNRT and CRT are the composites of human-intensive and computer-intensive interactions.

Figure 7 shows the relationship between work session time for 100 composite interactions and CRT for human-intensive interactions for both systems. For system A, work session times are 1624 seconds and 3391 seconds at CRTs of 0.5 second and 3.0 seconds, respectively, a 109 percent difference. On system B, for the same CRTs, work session times are 1154 seconds and 1922 seconds, a 67 percent difference.

At higher IUP, work session time is shorter. Shorter work session time results in shorter user session time only if the number of work sessions in a user session remains the same. That might mean that the user logs off after accomplishing his goal—after the successful execution of program A. This results in a smaller number of active users concurrently logged on the system but generating the same aggregate user work. On the other hand, if users terminate their sessions based on the time spent interacting with the system (for example, after one hour) then higher user productivity results in the users' accomplishing more work during

Figure 7 Work session time versus computer response time for both systems studied



user sessions, with no change in the number of concurrent logged-on users. This can occur, however, only if the computer has the power to execute the additional workload. Observations on system A indicate that in practice both conditions occur. A reduction in the CRT for human-intensive interactions by a factor of two resulted in fewer concurrent logged-on users for a period of a few days. This was followed by an increase in the number of concurrent users who generated a larger aggregate workload. Thus, the number of concurrent logged-on users—without considering CRTs, user productivity level, or aggregate user workload—is a poor measure of system service.

Cost

In the 1960s, the cost of doing work on a computing system was dominated by computer hardware costs. Moreover, jobs were submitted for processing in a batch mode. Therefore, batch

turnaround times affected project completion times. Technological advances since then have dramatically reduced the cost of computer hardware. With the shift from batch to interactive computing and with centralized systems concurrently supporting large numbers of interactive users, the cost of doing work on a computing system is dominated by the cost of the interactive user's time. User productivity during interactions with the computer system is one of several factors that now affect project completion times.

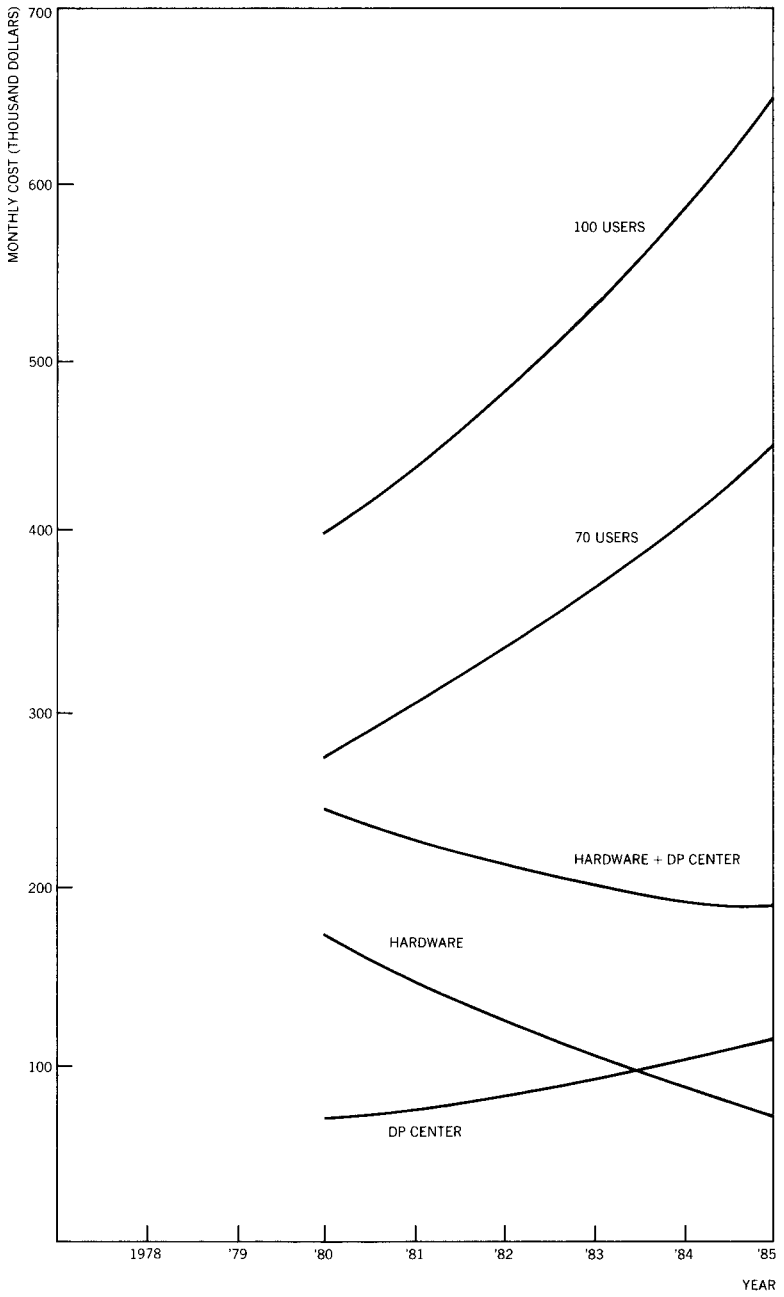
To examine cost differences in an installation, typical 1980 cost for user, hardware, and data processing center were assumed in this analysis. User costs, including salaries and benefits, were assumed to be four thousand dollars per month. The monthly lease cost for a typical System/370 Model 168 multiprocessor, including all peripheral equipment such as disks, tapes, printers, etc., was assumed to be 500 thousand dollars per month. The data processing center cost to support such a system was assumed to be 200 thousand dollars per month. Since interactive TSO work was done mainly during first shift and batch and other applications were processed during the second, third, and weekend shifts, we assumed that 35 percent of the hardware and data processing center cost was to be charged for interactive TSO, and 65 percent for batch and other applications. These costs are considered to be representative, although costs at particular data processing centers may differ widely.

For 1980, then, the cost of the aggregate user time on system B, with 70 active logged-on users, was 280 thousand dollars per month (i.e., 70 user months times four thousand dollars per user month). The assumed monthly hardware cost for interactive TSO was 175 thousand dollars (i.e., 35 percent of 500 thousand dollars), and the data processing center cost was 70 thousand dollars (i.e., 35 percent of 200 thousand dollars). In Figure 8, two curves are shown for the aggregate TSO user cost: the lower for 70 users (the average number of users on the system), and the upper for 100 users (the maximum number of concurrent users allowed on the system). Cost projections are made for 1981 to 1985, assuming that user costs increase at a rate of 10 percent per year, that system costs decrease at a rate of 15 percent per year[8] (reflecting technological advances), and that data processing center costs increase 10 percent per year.

The data show that the cost of the aggregate interactive user time is larger than the combined system and data processing center cost in 1980. By 1985, the user cost may be twice as large as the combined system and data processing center cost.

The average CRT for human-intensive interactions on system B is 0.84 second. Consider the alternative shown in Figure 7 of

Figure 8 TSO cost trends and projections for system B



operating at 0.25 second with a 36 percent reduction in work session time. In that case, the same user work would have yielded a potential savings of 100 thousand dollars per month (i.e., 36 percent of 280 thousand dollars) in the aggregate user cost

component for system B in 1980. On the other hand, operating at an average CRT of 3.0 seconds with a 44 percent increase in work session time would have resulted in a potential increase of 123 thousand dollars per month (44 percent of 280 thousand dollars). For the same application in 1985, using projected costs, the potential savings at 0.25 second would be 162 thousand dollars. The potential increase at 3.0 seconds would be 198 thousand dollars. These differences are in the user cost component only. The system cost to decrease response time when subtracted from the savings in user cost would yield a net savings to the installation.

trends in system management

In achieving economies of scale in large centralized computer facilities, system managers focus on maximizing central processor utilization and system throughput. Rarely are the processors allowed to be underutilized to ensure good computer response times to user interactions. In fact, the number of users concurrently allowed on the system is intentionally kept high to prevent the processors being idle for lack of work. A simple method of controlling poor CRT for the user while maximizing processor utilization has been to increase the number of interactive users until they complain of poor service, then reduce and stabilize the user population below the frustration level.

In configuring systems and selecting hardware, the primary emphasis has been on maximizing processor utilization. For example, high-performance I/O devices were installed to alleviate storage subsystem bottlenecks and to maximize processor utilization. Good system response times have been a secondary goal. These strategies and policies were appropriate in the early 1970s when system costs outweighed user costs. However, user cost now exceeds the combined hardware and data processing center cost, and the divergence is expected to continue in the 1980s. Therefore, new strategies and policies are needed that concentrate on system response time and user productivity and deemphasize processor utilization and system throughput.

On system B, the average human-intensive interaction executes 192K instructions and requests two I/O records to be transferred from disk storage. Supervisor instructions, paging I/O, and swap I/O requests are not included in this characterization, which is summarized in Table 1. Human-intensive interactions are swapped into main memory at least once, and the number of page fault I/Os is a function of main memory contention.

Components contributing to CRT are instruction execution time, disk file I/O and paging I/O times, swap I/O time, and delay times waiting on queues for these and other system resources. Of these components, the instruction execution time for human-intensive interactions—including supervisor instructions (estimated as an additional 40 percent) on the system B processor (System/370

Table 1 System B workload characteristics

Characteristic	*Human intensive*	*Computer intensive*
Interactions	96.5%	3.5%
CRT seconds	0.84	36
Instructions per interaction	192K	6.8M
I/O per interaction	2	106

Model 168)—is approximately 0.1 second. The rest of the CRT is due to disk file I/O, swap and paging I/O times, and processor and I/O queue delays. Therefore, high-performance I/O devices and larger main memory buffers that avoid I/Os altogether, along with system scheduling policies, play a dominant role in reducing CRTs to the subsecond range for human-intensive interactions. For computer-intensive interactions, with larger numbers of both instructions executed and I/O requests, both faster processors and high-performance I/O devices are important in reducing CRTs.

Concluding remarks

The cost of accomplishing work on a central computer facility—providing service to a large number of interactive users—has shifted, with the cost of the aggregate user time being the dominant component. Moreover, user costs are expected to continue to increase and system costs are expected to continue to decrease in the coming decade. This suggests that computer systems should be managed for maximum user effectiveness rather than for maximum machine usage.

The data show that on the two systems analyzed, interactive user productivity and user response time are correlated with computer response time. The slopes of the curves are significantly larger in the 0.25- to 1.0-second computer response time (CRT) range than for values of CRT greater than 1.0 second. That they are correlated does not, however, imply that higher CRT causes higher user response time (URT) and lower interactive user performance (IUP). Changes in user work pattern may cause higher URT at higher CRT. Unfortunately, controlling the environment to isolate and investigate such factors in production systems is not possible. We did, however, examine other system measures; they suggest that users do not significantly modify their work pattern in the 0- to 3-second range.

Assuming that persons are more effective at short CRT, the savings in user cost was shown to be significant, particularly for subsecond CRT. Since network and transmission delays are quite large for remote users, locally attached terminals may be the preferred alternative in meeting subsecond SRT. The effect of variation in user work pattern on user productivity is among the issues that need further investigation.

ACKNOWLEDGMENTS

I thank S. Goldstein and C. T. Apple for their enthusiastic support and for the many discussions that helped formulate several of the concepts and methodology. My thanks also go to A. C. Munce and G. R. Henry for their assistance during the analysis and their review and critique of this paper. And I wish to thank P. R. Conrad for making the data available to me and D. B. Edwards for his support of this work.

CITED REFERENCES

1. W. J. Doherty and R. P. Kelisky, "Managing VM/CMS systems for user effectiveness," *IBM Systems Journal* **18,** No. 1, 143–163 (1979).
2. S. J. Boies, "User behavior on an interactive computer system," *IBM Systems Journal* **13,** No. 1, 1–18 (1974).
3. *OS/VS2 TSO Terminal Users Guide*, GC28-0645-4; available through IBM branch offices.
4. *Advanced Communications Facility for VTAM, General Information Introduction*, GC27-0462-2; available through IBM branch offices.
5. *OS/VS2 Resource Management Facility, General Information Manual*, GC28-0921-2; available through IBM branch offices.
6. R. B. Miller, "Response time in man-computer conversational transactions," *AFIPS Conference Proceedings, Spring Joint Computer Conference* **33,** 267–277 (1968).
7. M. Grossberg, R. A. Wiesen, and D. B. Yntema, "Experiment on problem solving with delayed computer responses," *IEEE Transactions on Systems, Man, and Cybernetics* **SMC-6,** No. 3, 219–222 (March 1976).
8. L. M. Branscomb, "Computing and communications—A perspective of the evolving environment," *IBM Systems Journal* **18,** No. 2, 189–201 (1979).

The author is located at the IBM General Products Division, 5600 Cottle Road, V41/98, San Jose, CA 95193.

Reprint Order No. G321-5156.