SIMULATION OF THE TIME-VARYING LOAD ON FUTURE REMOTE-ACCESS IMMEDIATE-RESPONSE COMPUTER SYSTEMS

H. A. Anderson, Jr.
IBM Watson Research Center
Yorktown Heights, New York

Abstract

A GPSS/360 program has been developed to model the behavior of a remote-access immediate-response computer system serving many concurrent users. Inputs to the model include characteristics of the users, the application category, the communication equipment, and the computer system. The model uses this information to determine the time-varying load imposed on the system in giving each on-line user immediate response.

1, INTRODUCTION

By the mid-1970's there is expected to be a significant number of remote-access immediate-response computing (RAIR) systems serving from 500 to 2500 concurrent users. The future technological development of integrated circuitry, communication facilities, and programming should make this type of system economically feasible and justifiable for large organizations. The increasing sophistication of computer usage and the increasing number of computer users will also serve to accelerate these developments. At this point in time, a thorough system analysis must be conducted during the early phases of the design of a RAIR system in order to end up with an efficient and economical system. Realistic performance criteria for the different RAIR subsystems have to be established so that the appropriate design tradeoffs can be made. Simulation is the best system analysis methodology for establishing these performance criteria. This type of system is very complex and difficult to formulate analytically due to the enormous number of states required to represent all the possible states of the system. One is neither interested in a worst-case deterministic analysis which can lead to an expensive over-design, nor in a mean-value analysis which in the "random" real world would be meaningless. Unfortunately, analytical techniques for any detailed computer system modeling tend to be mean-value analyses.

The message traffic arriving at a RAIR system is a random process. The demands placed on the system's resources are represented as random variables. It is important to obtain estimates of the probability distributions which represent the demands of the message traffic on the system resources. For a computer system developer it is equally as important to determine how these probability distributions vary from application area to application area. This kind of information may be obtained from simulating a RAIR system. The resource demands of the message traffic (the workload) can then be translated into performance criteria for the different components of the RAIR subsystems.

2. CONSIDERATIONS IN SIMULATING A FUTURE RAIR SYSTEM

A typical RAIR system is shown in Fig. 1. It is unrealistic and impractical to begin the simulation of a proposed large RAIR system of the middle 1970's at a high level of functional detail. To do so would require estimates of future hardware and software technological developments that are unrealistically detailed. The computing time spent in simulating would be excessive. The results would be no more accurate and believable for a detailed model than for a less-specified model since the former would require many more parameters to be estimated. The focus of the GPSS/360 model described in this paper is on characterizing both the application areas to be serviced by a RAIR system and on the behavior of users at their terminals. User reaction times (input insertion time, reading time, think time) upon getting immediate response to their queries determine the workload to be serviced by the system. The GPSS/360 model requires information on user characteristics, application area characteristics, communication subsystem characteristics, and computer subsystem characteristics in order to conduct a simulation experiment (Fig. 2).

It has been observed in our simulation experiments (and for some real systems) that a moderately loaded RAIR system behaves as a linear stochastic system in which the workload increases linearly with increasing message traffic while the response time remains stationary. Response time is a probabilistic phenomenon. Its being stationary means that its statistical properties are not appreciably changing due to an increasing workload. Our simulation model of a RAIR system is "tuned" by repeated experimentation to behave as a linear stochastic system. The response-time probability distribution we have observed is skewed reflecting the effects of the queueing that occur in each subsystem. Stabilizing the response time constrains the system to process its instantaneous workload within some upper bound on time. This is a main factor in establishing the performance criteria required. Our objective is not to simulate the exact instant, for example, when a data base access is to be enacted or a program overlaid, but instead to determine the number of data base accesses or program overlays that occur during a given time interval. The real time clock of our RAIR simulation experiments resolves time to milliseconds. If a simulation experiment is conducted where a RAIR subsystem component becomes heavily utilized (75% to 85% depending on the variation of the subsystem service time) the overall system behavior departs from linearity for an increasing workload and the mean response time increases exponentially. The performance parameters of the overloaded subsystem component have to be adjusted to bring the response time within bounds again so that the results of this experiment can be compared with the results of previous experiments.

There are many RAIR systems operational today in different application areas (demand deposits, consumer credit, customer information, travel reservation, etc.). The logical organizations of these systems are similar. A generic flow diagram has been abstracted to represent the possible processing paths a message may take through such a system. The flow diagram and the messages (translations) flowing through it have also been parameterized to model the system resource demands of the messages as they pass through the system. The different system resource demands of the message traffic are tabulated during the course of a simulation experiment. This generic flow diagram (model) of a RAIR system coupled with a model of the behavior of users sitting at terminals submitting messages into the system represents the GPSS/360 model that was developed to analyze the system requirements of future RAIR systems.

There is not much known or published about user behavior at different types of terminals, the performance of present-day RAIR systems, the future RAIR system requirements of different application areas, nor the impact of developing technologies on RAIR subsystems. Despite this lack of knowledge, the demands are already present for the development of these systems. System analysis can still proceed to assist the designer even if it is initially limited to a sensitivity analysis. An understanding of how best to distribute system response time among the different subsystems involved in processing a user's message can be obtained through simulation. Intensive efforts will have to be undertaken by all those concerned with future RAIR systems to obtain information on system requirements so that more detailed and realistic simulation models may be developed as we proceed into the future.

We expect to proceed iteratively through design and analysis of RAIR systems, starting with our generic model to perform sensitivity analyses. Our first goal is to determine what subsystem components will be the performance "bottlenecks" (and under what conditions) in the RAIR systems of the future. Having satisfied this goal, design work can then focus on eliminating these performance problems. In addition, we will have also determined the values of the input parameters for which more detailed simulation experimentation will be conducted using modified RAIR models reflecting specific system organizations and encompassing new or expanded functional requirements.

3. CONSIDERATIONS IN SIMULATING THE USERS OF A RAIR SYSTEM

A user session at a terminal can be described as occurring in phases. The typical phases are log-on, enter messages, and log-off. During each phase a user's behavior, his reaction times, and the demands his messages place on the system resources are different. With the number of concurrent users averaging in the hundreds to possibly in the thousands, there would be a significant number of people in each phase. This would make the population of concurrent users appear heterogeneous to the RAIR computer system even if the system was being used to service a single application area.

A user in a particular phase of work at a terminal can be found in either one of three states: the input insertion state (inputting a message), the wait state, or the think state (see Fig. 3). In some application areas where response time is expected to be 5 sec or less for 99% of the inquiries, approximately 90% to 95% of the users would be found in either the input state or the think state at any one time. For this type of system the workload is derived from 5% to 10% of the concurrent users of the system. The reason for this is that human reaction times in typing a message or inserting a card in a reader, reading a response and then thinking about it would be longer on the average than the 2 to 5 second response time desired of a RAIR system.

In modeling the users' behavior it is important to account for novice users and for experienced users. Their reaction times, their overall terminal activity, and their demands on system resources would be significantly different. A novice user is likely to be instructed in the use of the system through computer-assisted instruction procedures while he is entering his messages. An experienced user would be finding and demanding procedures which would permit him to reduce the amount of work he has to do in entering and responding to routine messages.

Over the course of time, the population of users of a RAIR system will evolve from a majority of novices to a majority of experts. Thus the workload on the system will also change. This sort of evolution has to be accounted for in any system analysis of a RAIR system. The GPSS/360 model permits a user to input either a single segment or a multiple segment message (see Fig. 4). System responses can either be a single frame or a multiple frame message (video displays are assumed). It is felt that this capability would permit a user to be modeled either as a novice (single segment input message, multiple frame response) or as an expert (multiple segment input message, single or multiple frame responses).

4. THE CHARACTERISTICS OF A USER OF A RAIR SYSTEM

In GPSS/360 parlance, a user is represented by a transaction being sequenced through the model flow diagram. The length of time a transaction is active in the model represents the user's holding time of a terminal. When a user has completed inserting a message into the system, the transaction representing him leaves the user section of the model and then enters and is sequenced through the section of the model representing the RAIR system. The parameters which characterize a user and have to be supplied (typically as probability distributions) to the model are:

- arrival rate of users to the system
- indication of whether a user is to input (respond to) single segment (frame) or multiple segment (frame) messages
- input rate, i.e., typing rate, etc.
- think time of a user
- reading rate of a user
- reaction times between different phases of a terminal session
- balking probability of a user when he is inputting a message causing him to repeat
- probability of user having an error in a message.

5. THE CHARACTERISTICS OF AN APPLICATION AREA TO BE SERVICED BY A RAIR SYSTEM

There are many different application areas which are already using RAIR systems. The workload on a RAIR system varies widely from application to application. A document processing and retrieval system serves users who can make many different types of requests for service. Their dialogue with the computer can be quite varied. A travel reservation system serves many users enacting similar business transactions at their terminals using very well defined scripts to cover their dialogue with the system. However, both systems have the common attribute of giving immediate response to a query. The parameters which assist in characterizing an application area and have to be supplied to the model are:

- number of messages to be entered during a user session
- log-on and log-off procedures
- number of segments to an input message
- number of frames to a response
- number of characters to an input message segment
- number of characters to an output message frame
- number of data bases to be accessed by a message

- number of accesses to the directory of a data base
- number of accesses to a data base
- number of I/O characters transferred per data base access
- number of program accesses required to process an input message and to prepare a response
- number of I/O characters transferred per program access
- amount of workspace required to support a user session
- probability of a user inputting multiple segment messages
- probability of a response to a user being a multiple frame message
- number of data bases that can be accessed and their probabilities of being accessed
- probability of an update message seizing a data base while updating a record in the data base
- probabilities of a message being an update message or a query message
- I/O activity for system audit.

6. THE CHARACTERISTICS OF THE RAIR COMMUNICATION SUBSYSTEM

The type of terminals users of a RAIR system have directly impacts the instantaneous workload on a system. A typewriter terminal with an output rate of 15 to 20 characters/sec can definitely slow a user down, while a 1000 character video display can speed up and simplify the dialogue between the man and the computer. Remote multiplexors are becoming standard communication equipment, which are used to increase the utilization and throughput of a communication line (channel). Since a communication line is a shared facility when a remote multiplexor is used, messages can be delayed in the communication subsystem due to queueing and multiplexor processing. Line control procedures, transmission speeds, message routing, mode of transmission (half-duplex or full-duplex), and line error rates all affect the message processing time of the RAIR communication subsystem. The processing time in this subsystem can be as much as 1 to 1.5 seconds. If a response of 2 to 5 seconds is required for satisfactory system performance, then the communication subsystem can appreciably impact the performance of a RAIR system by requiring the computer subsystem to process a message under tighter time constraints. The parameters which assist in characterizing the RAIR communication subsystem and have to be supplied to the model are:

- maximum terminal rate for inputting and outputting messages
- number of terminals on a communication line
- number of remote multiplexors on a communication line
- communication line transmission speed
- mode of transmission: half-duplex or full-duplex
- communication line control procedures
- communication line error rates requiring message retransmission
- remote multiplexor message processing delays
- message priority assignment
- message routing procedures
- message header length for identification.

7. THE CHARACTERISTICS OF THE RAIR COMPUTER SUBSYSTEM

The attempt in this model is to represent as little specific computer hardware as possible, but instead to concentrate on basic time-consuming functions and modeling all the situations where queueing and processing delays would be incurred by a message as it is sequenced through a RAIR computer system. The main components of this subsystem are the teleprocessing component, language translation component, data management component, error diagnosing component, response preparation and system audit components, and system accounting component. Message processing times have to be estimated for each component. The number of phases of processing for a message in each component is dependent upon the application area (see Fig. 6). The parameters which assist in characterizing the RAIR computer subsystem and have to be supplied to the model are:

- message processing times for each component in milliseconds
- queueing and scheduling strategies
- direct access storage device access times
- configuration of channels for input and output

8. THE PERFORMANCE PARAMETERS MEASURED DURING A RAIR SIMULATION EXPERIMENT

The following parameters are sampled at the end of every half second, one second, or at different points in a message's path through the flow diagram of the model:

- response time
- user terminal holding time
- number of users on the system
- number of users in think state, wait state, and input state
- number of messages and characters inputted (outputted) per second
- number of data base accesses, program requests, message queueing accesses, and error-diagnosing accesses per second
- utilization of the communication lines, multiplexors, processors, and I/O channels
- number of interrupts services per second
- number of characters transferred over the I/O channels per second
- input (output) message lengths

At the termination of the simulation experiment, these parameters are outputted as probability distributions along with their mean values and variances. The model also gathers time series information on most of the above parameters over one-minute time intervals of steady-state system operation. These time series give added information and insight into the dynamics of the behavior of a RAIR system. It is not unusual to observe the number of concurrent users on such a system to change by over 10% during a one-minute interval. Analysis of these time series such as autocorrelation, cross-correlation, and spectral analysis should contribute significantly to a better understanding of the behavior of these systems and lead to the development of new control programs for future systems.

9. THE DESCRIPTION OF THE RAIR SYSTEM MODEL

Figure 5 represents an overview of the steps involved in processing a message within a RAIR system. The RAIR computer subsystem is modeled in greater detail than the communication subsystem (see Fig. 6). The components of the computer subsystem which are depicted will require significant computing capability to service the message traffic of the future. Furthermore, they will be the significant contributors to the overall I/O-related activity within the computer subsystem. For these reasons they have been represented in the model.

The section of the model that represents the communication subsystem models the behavior of the terminals, multiplexors, and communication lines in the RAIR system. The impact of changing the values (or probability distributions) of the different communication subsystem performance parameters on the total system's performance may be investigated, i.e., increased line speeds, or more terminals on a line, etc. This permits total system design trade-offs to be made. In addition, the impact of new technologies for communication can be evaluated in the context of the communication requirements of RAIR systems.

The language translation component of the computer subsystem processes a message in several phases. Programs like the General Information System (GIS) and the Information Management System (IMS) presently contain their own language translators. The desire is to get the user to express his messages in a language more English-like in its structure and syntax. This would permit the user to work more effectively. In addition, he should not be concerned about the internal format and organization, and physical location of the data that he wants to access. This trend should continue in the future, as the attempt now is to get more and more people to conduct their work at terminals. This generalized approach to language translation for RAIR systems will place great computational demands on future computer processors.

The model of the data management component has the capability to process messages requiring accesses to a single data base or multiple data bases. The approach is to represent the amount of computer processing required in data base searching, and in processing the results of a search. The overall I/O activity required to perform these functions is also modeled. Additional delays can be incurred by a message in the data management component due to a message requiring access to a data base which another message has already seized and is updating. The frequency with which this situation occurs has direct impact on response time and thus has to be included in the simulation model.

The other components of the RAIR computer subsystem that are modeled will all be moderately utilized by the message traffic. The system accounting component is used to process messages representing user requests to log-on or log-off the system. This may occur as much as 10 times a second on the average for some application areas. The error-diagnosing component primarily is involved in determining what diagnostic message should be transmitted back to a user who has submitted an incorrect message. All errors would also be recorded. The response preparation and system audit component sorts, edits, and formats the information to be presented to the user as a response to his message. It also records all changes to the data bases made by the update messages and delete messages inputted to the system. The I/O activity of these components is accounted for in the simulation model.

10. AN EXAMPLE IN USING THE RAIR SIMULATION MODEL

The example discussed in this section covers two cases: case 1 is where 90% of the users are novices and 10% are experts, and case 2, where 10% of the users are novices and 90% are experts. Table 1 summarizes the most significant input parameters used in the simulation of the two cases. Simulation experiments in both cases were conducted for the following arrival rates of users (assuming the arrival process is a Poisson process) 1, 1.5, 2, 2.5, and 3 users/second. Table 2 summarizes some of the output statistics for the experiments where the user arrival rate was 3 users/second.

The computer subsystem modeled for this example contains three processors (facilities): a teleprocessing processor, a message processor, and data management processor. The graphs in exhibit 1 show the utilization of these processors as a function of the arrival rate of users for the two cases. The data management processor has a utilization of 84% in both cases at an arrival rate of 3 users/second. The arrival rate of 3 users/second can be translated in terms of the average number of users on the system. For case 1 it is 929 users and for case 2 it is 775 users. Exhibit 2 depicts how the response time probability distribution changes as the arrival rate of users increases. The degree of change is not very much and probably would not be perceivable by the users. At an arrival rate of 3 users/second in both the cases, the RAIR system's behavior is starting to depart from a linear response to an increasing workload. The high utilization of the data management processor is beginning to affect system response time. A utilization of a subsystem component of over 85% usually increases response time noticeably. The reason that in this example the system response time degradation is not as bad as one might expect is that the variance of the probability distributions of the service times of the processors is not large. The variation in the service time of a queueing facility definitely limits how high the utilization of the facility may be before the average time spent waiting by a customer for service gets unacceptably long. The RAIR simulation model gives us the ability to determine the workload where a given RAIR system's performance begins to depart from a linear system behavior and to see the effect this has on system response time. Exhibit 3 was included to illustrate the dynamic variation of the instantaneous workload of the computer over a 1-minute time period when there are hundreds of concurrent users on a RAIR system.

The holding time probability distributions for the two cases (exhibit 4) illustrate what one would expect, and that is, that the holding times for a population of users where 90% are experts would be shorter than the holding times of a population of users where, instead, 90% are novices, trying to accomplish the same work. The remaining exhibit (exhibit 5) presents graphs that depict the linear behavior of the mean values of certain of the system performance parameters as a function of the arrival rate of users or mean number of users on the system. Limiting our remarks to making only general observations about the results of the simulation experiments for this example, the following may be stated.

- the rate of increase of the number of users in the wait state (a measure of the workload on the system) increases much more slowly with increasing arrival rate of users than does the total number of users on the system (most users spend a majority of their holding time either in the think state or in the input insertion state)
- holding time statistics directly determine the number of people on the system for a given arrival rate of users and is a function of what people are trying to accomplish at their terminals
- the demands placed on a RAIR communication subsystem can be significantly different depending on whether primarily novices or experts are using the system
- experts create a significantly greater workload on a RAIR computer subsystem than novices do (on a per-user basis)
- for high message input rates, quick-access direct access storage devices or even large capacity storage may have to be used for message queueing
- the number of interrupts to be serviced

> per second goes up quite rapidly with increasing message traffic and could contribute to excessive system overhead unless these interrupts are efficiently serviced.

The statistical estimation problem is significant in simulating a RAIR system and typical of a system containing queues. The samples for a given parameter are autocorrelated. This problem becomes increasingly severe and costly in terms of computer time as the utilization of any one major RAIR subsystem component gets above 75%. The analyst involved in simulating a RAIR system has to take this into account in determining how long a time period of system operation should be simulated in order to obtain reasonable estimates of the statistics of the system performance parameters. We have been using a 360/91 computer to simulate time periods of a half-hour of RAIR system operation. If the number of users on the system is under 900, the elapsed real time on the 360/91 computer is less than the simulated time for the RAIR model. There are other statistical estimation problems which come into play in experimenting with a RAIR simulation model and these have to deal with determining when the model has reached steady-state behavior and in estimating the tails of the probability distributions of the system performance parameters. These problems will not be discussed in this paper as there are no general theories to guide a systems analyst in coping with them.

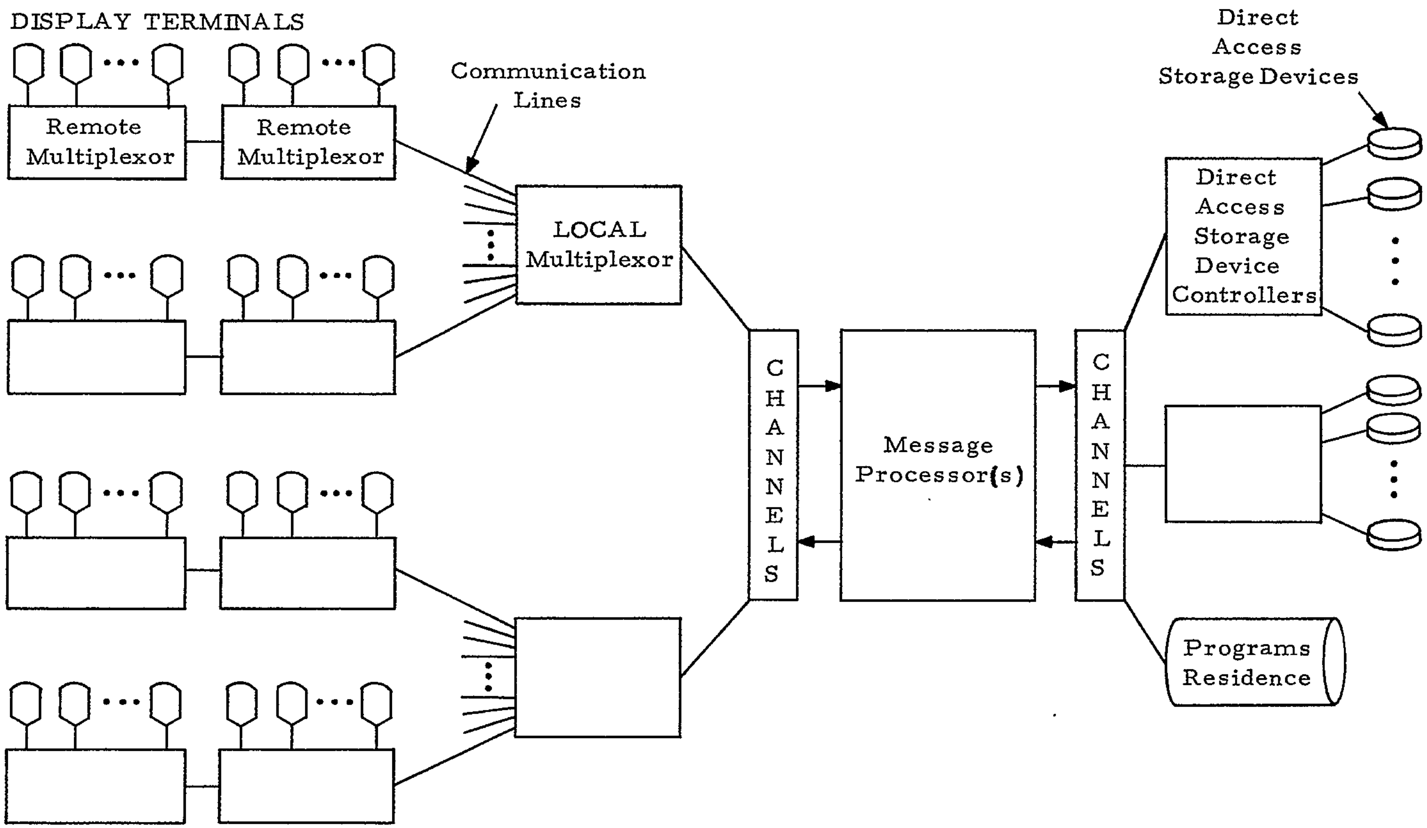
11. CONCLUSIONS

Simulation is the best methodology in performing the system analysis required to determine the performance criteria of the RAIR systems of the mid-1970's. The logical organization of the information flow in a RAIR system is very similar for many different application areas; thus, a generic simulation model can be developed to represent the behavior of such a system and its users. The system response time reflects the level of activity in both the communication subsystem and in the computer subsystem. Therefore, a total system approach to modeling has to be done initially at the expense of the level of detail of the model. A total system model can be used to determine the region in the input-parameter space where future simulation experiments using more detailed RAIR system models should be conducted.

The future computing and communication requirements of the different application areas where RAIR systems will be used will have to be determined. All those involved in either the design, implementation, or use of these systems will have to work together in obtaining this information. Realistic simulation experiments can then be conducted to support the system design efforts in making system design trade-offs. The manpower, money, and time that will be committed to implementing, installing, and operating RAIR systems will be considerable and an expensive worst-case approach to system design cannot be conducted.

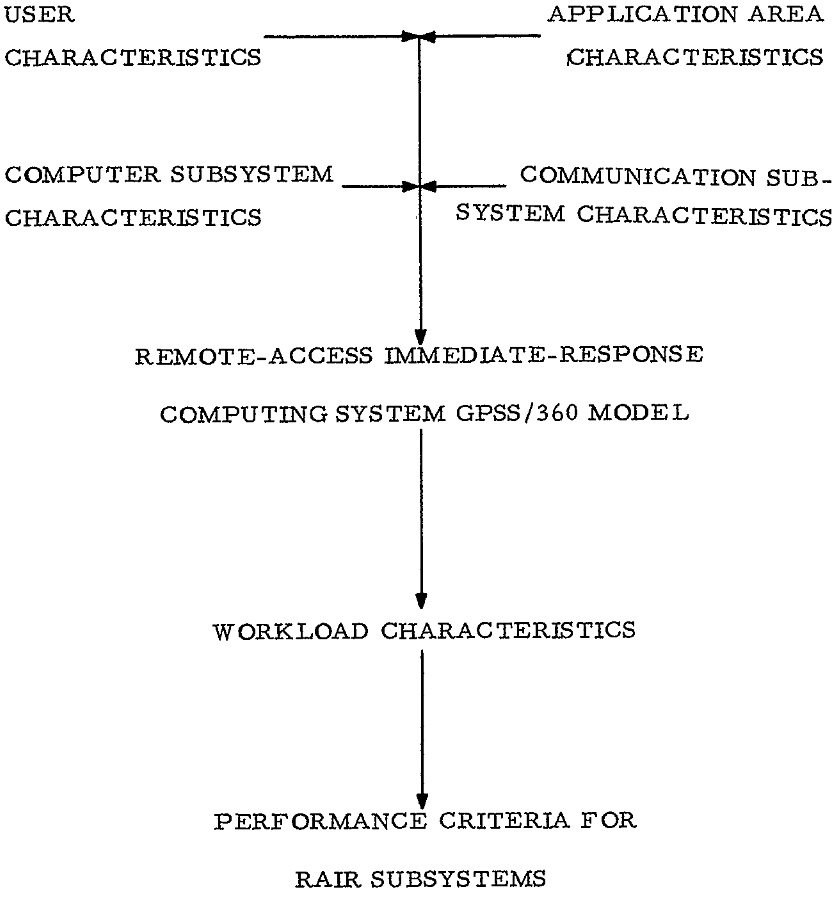
12. ACKNOWLEDGMENT

I appreciate all the comments and suggestions I have received from W. A. Notz, who encouraged and supported the development of the RAIR system simulation model.



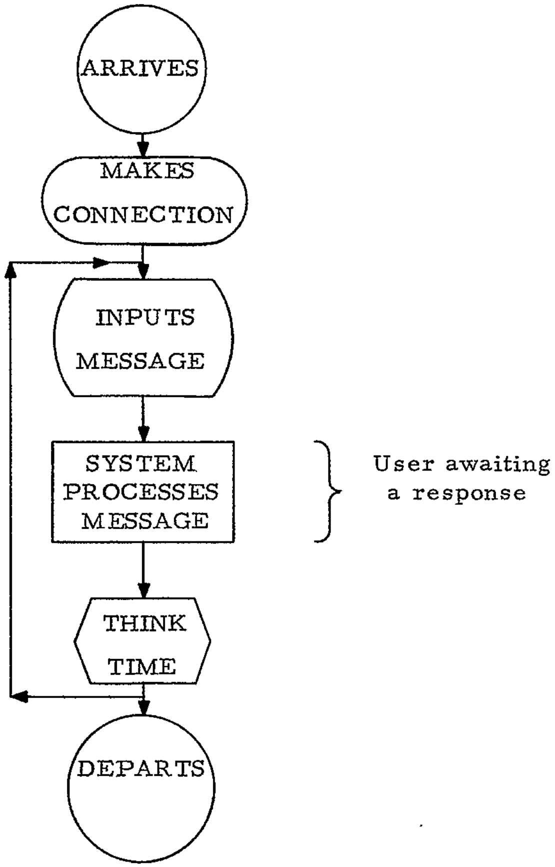
A Typical Remote-Access Immediate-Response System

FIGURE 1



Information Requirements of the RAIR System Simulation Model

FIGURE 2



Overview of a User Session

FIGURE 3

Detailed Model of a User Session

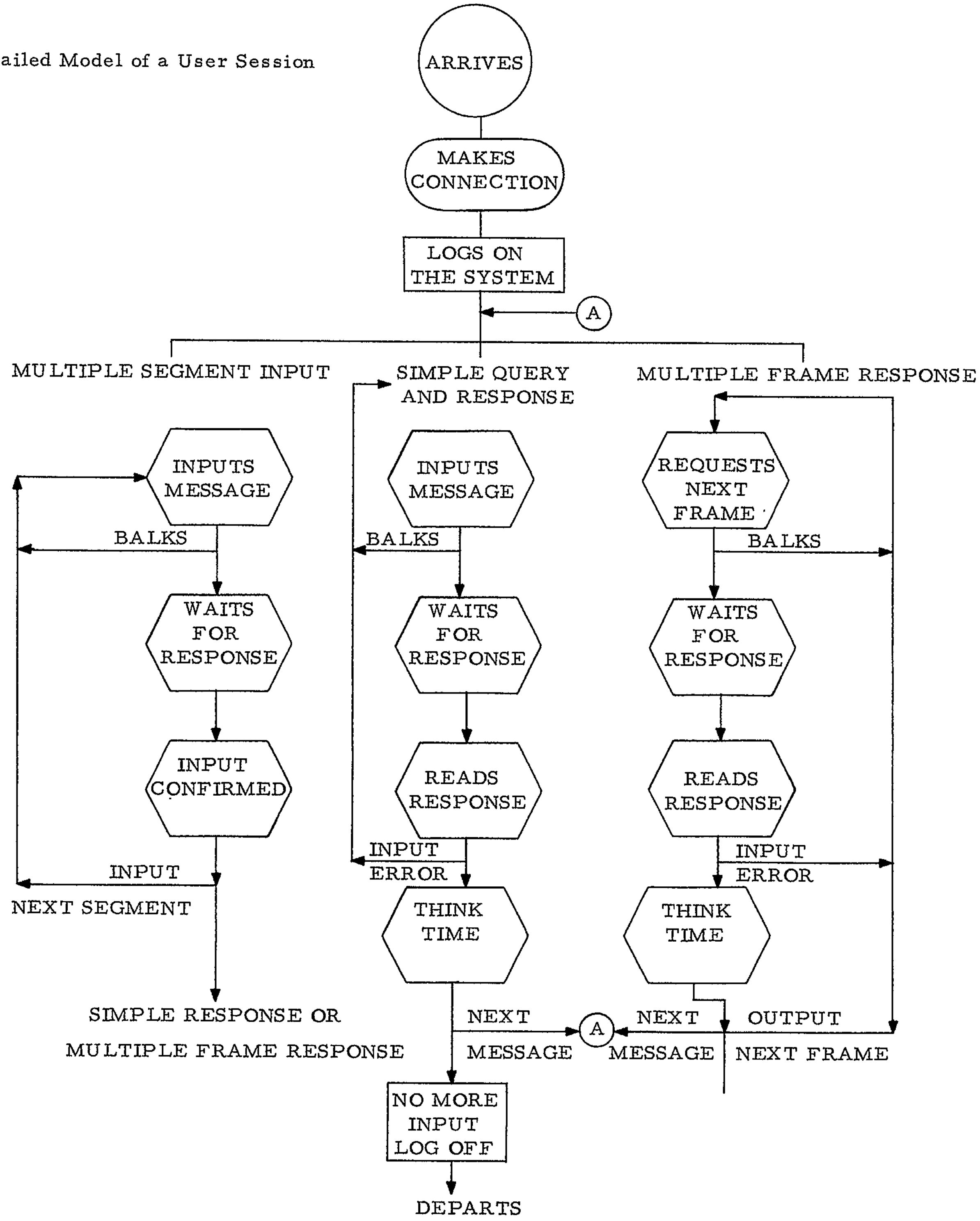
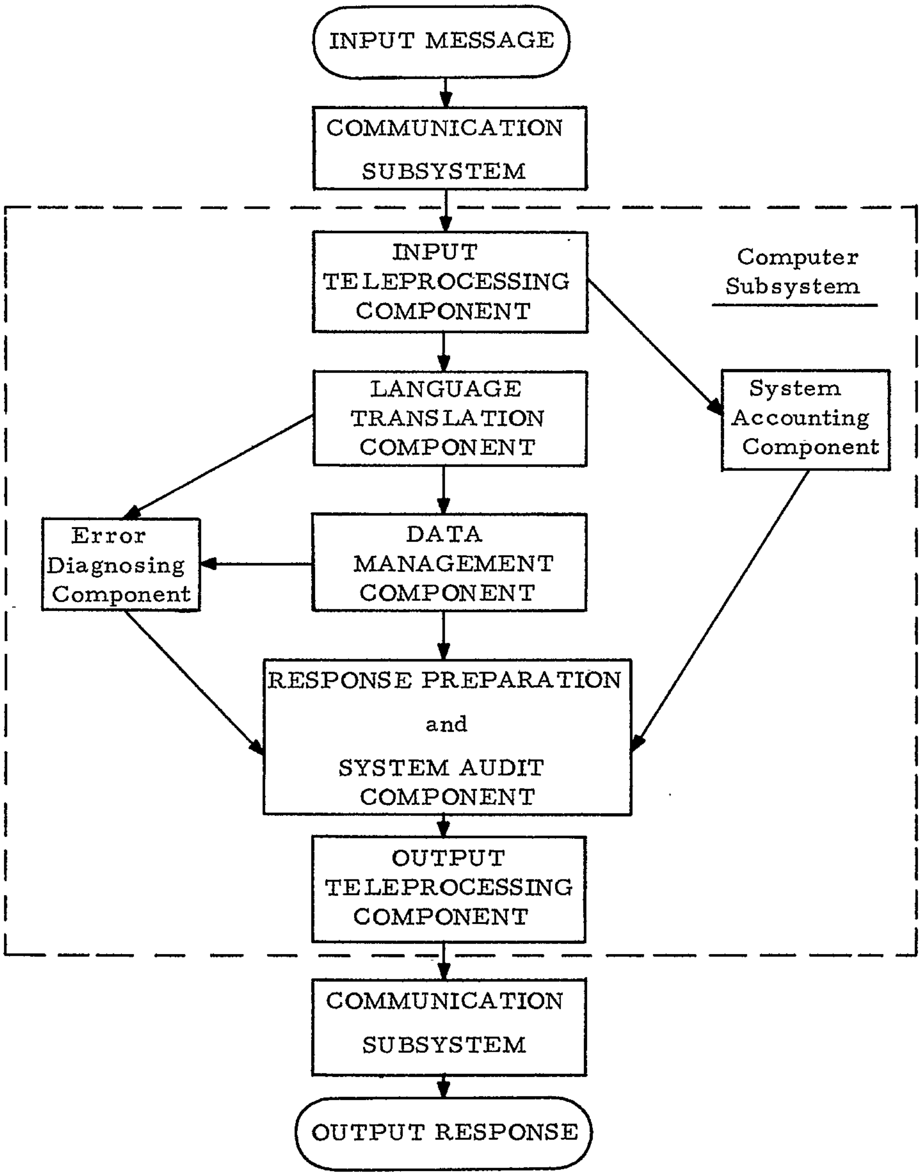


FIGURE 4



Overview of the Message Flow in a RAIR System

FIGURE 5

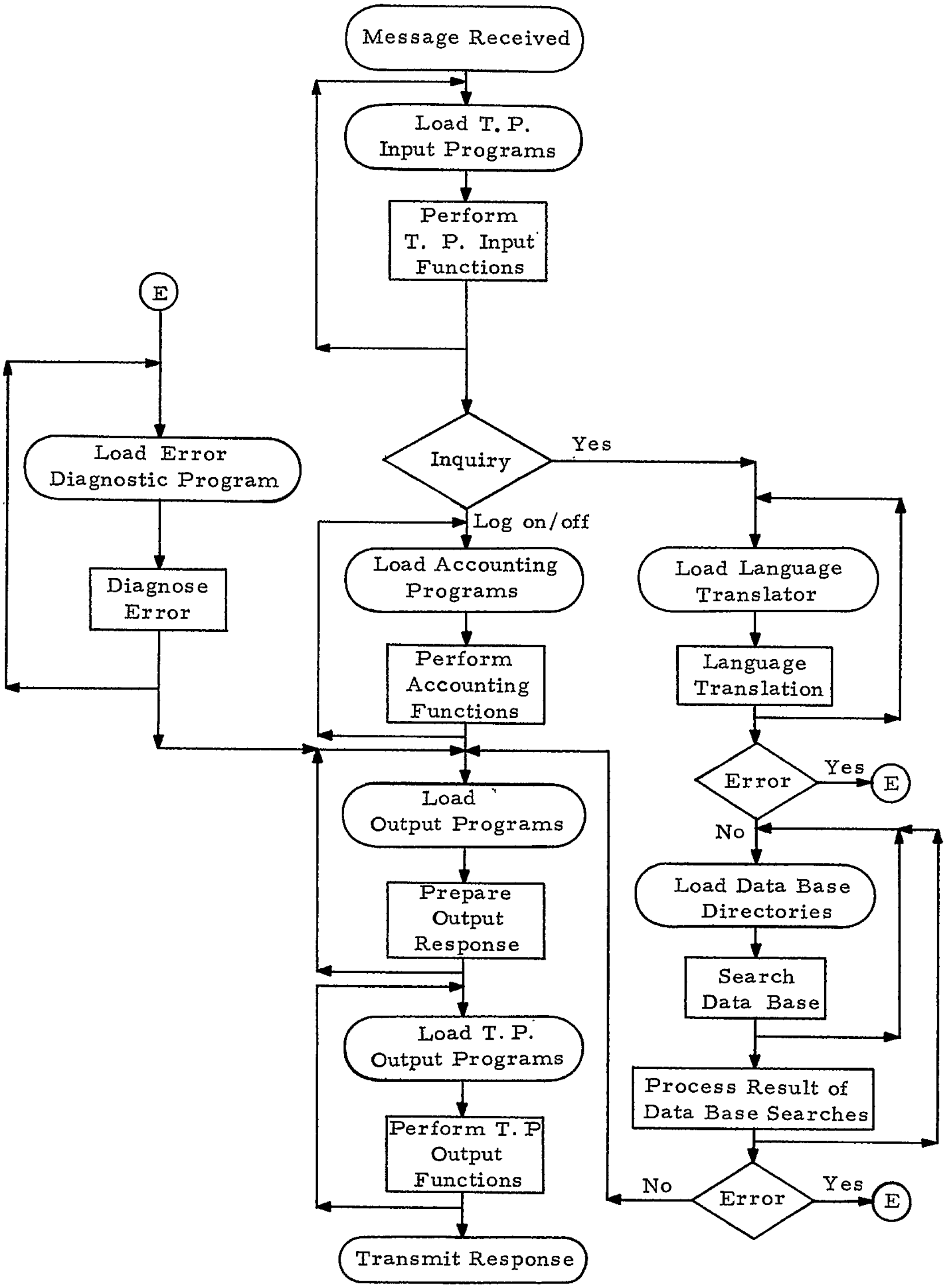


FIGURE 6. Detailed Model of the Message Flow in a RAIR System

Input Parameters	Novice	Expert
Mean Think Time	1.5 to 3 sec	0.5 to 1.9 sec
Typing Rate	1.5 to 3 char/sec	1.5 to 3 char/sec
Reading Rate	12 to 25 char/sec	12 to 25 char/sec
Input Error Probability	0.001 to 0.05	0.001 to 0.05
No. of Messages	2 to 3	4 to 6
No. of Input Message Segments	1	2 to 3
No. of Frames of Output	2 to 4	1 to 2
Input Message Segment Length	1 to 25 char.	30 to 60 char.
Output Frame in Characters	150 to 250 char.	200 to 400 char.
Number of Data Bases Accessed	1 to 2	2 to 4
Directory Accesses/Data Base Accessed	1 to 9	1 to 9
Accessed/Data Base Accessed	1 to 4	1 to 4
Probability of an Update Message	0.25	0.25
Probability of an Inquiry Message	0.75	0.75
Number of Terminals/Remote Multiplexor	10	10
No. of Multiplexors/Line	1	1
Input Line Speed	2400 bps	2400 bps
Output Line Speed	4800 bps	4800 bps
No. of Lines	100	100
No. of Phases of Language Translation	4	4
Computing Time/Phase	2 to 6 ms	2 to 6 ms
Computing Time/Data Base Search	4 to 10 ms	4 to 10 ms
Computing Time to Edit Results of Searches	2 to 6 ms	2 to 6 ms
Number of Phases of Response Preparation	2	2
Computing Time/Phase	2 to 6 ms	2 to 6 ms
No. of Phases of System Accounting	2	2
Computing Time/Phase	1 to 3 ms	1 to 3 ms
Computing Time for TP Input Processing	2 to 8 ms	2 to 8 ms
Computing Time for TP Output Processing	1 to 9 ms	1 to 9 ms

TABLE 1.

Output Parameters	Case 1	Case 2
Holding Time (sec)	311 (510)	258 (420)
Response Time (sec)	1.58 (5.00)	1.6 (6.50)
Users on the system	929 (980)	775 (840)
Users waiting for a Response	69 (95)	49 (75)
Users in the Think State	545 (600)	262 (300)
Users in the Input State	315 (360)	464 (540)
Messages Submitted/sec	44 (56)	31 (44)
Input Characters Transmitted/sec	1100 (1600)	1030 (1600)
Output Characters Transmitted/sec	8200 (12,000)	4200 (7000)
I/O Channel Characters Transferred/sec	1,400,000 (2,000,000)	1,400,000 (1,800,000)
Program Requests/sec	159 (200)	118 (170)

(the number within the parentheses is the value which 99% of all readings fall below; the other number is the mean value)

TABLE 2

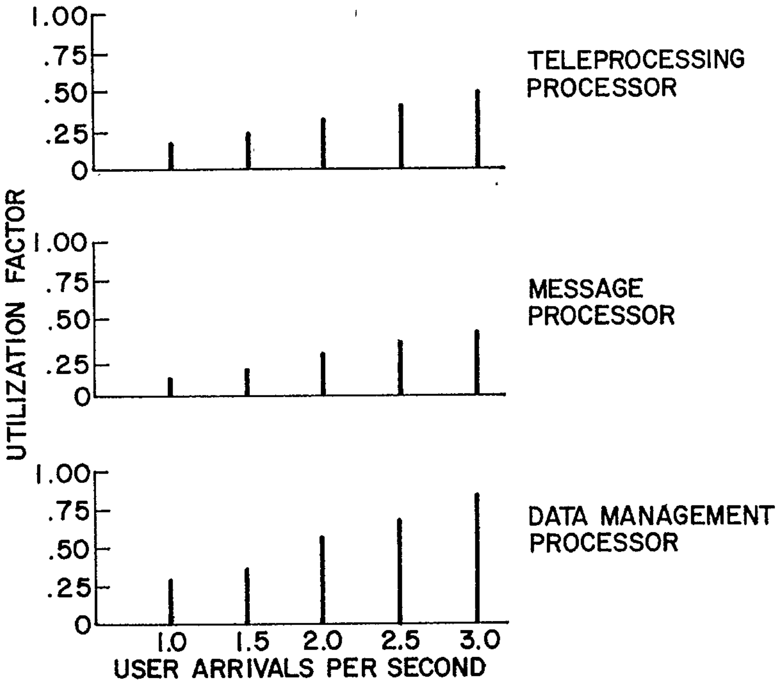


Exhibit 1A. Component Utilization as a Function of User Arrival Rate (Case 1)

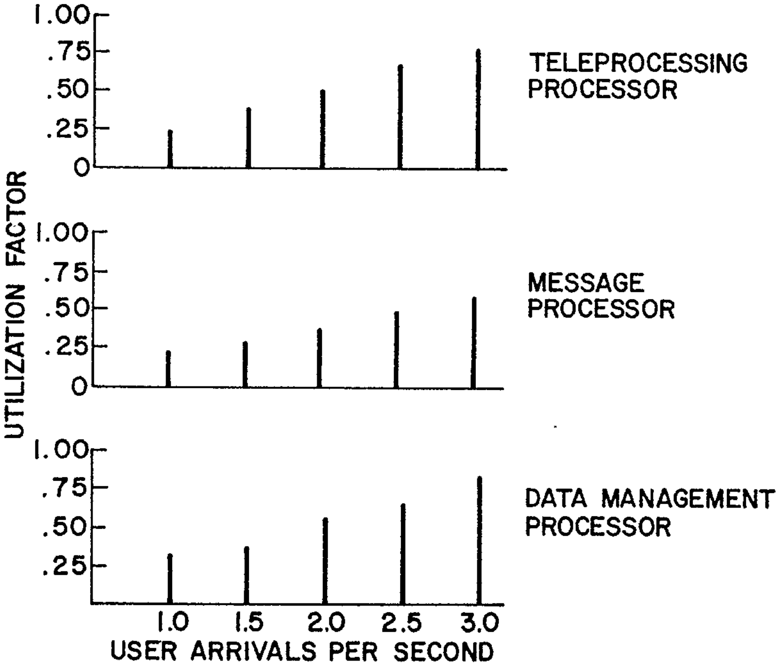


Exhibit 1B. Component Utilization as a Function of User Arrival Rate (Case 2)

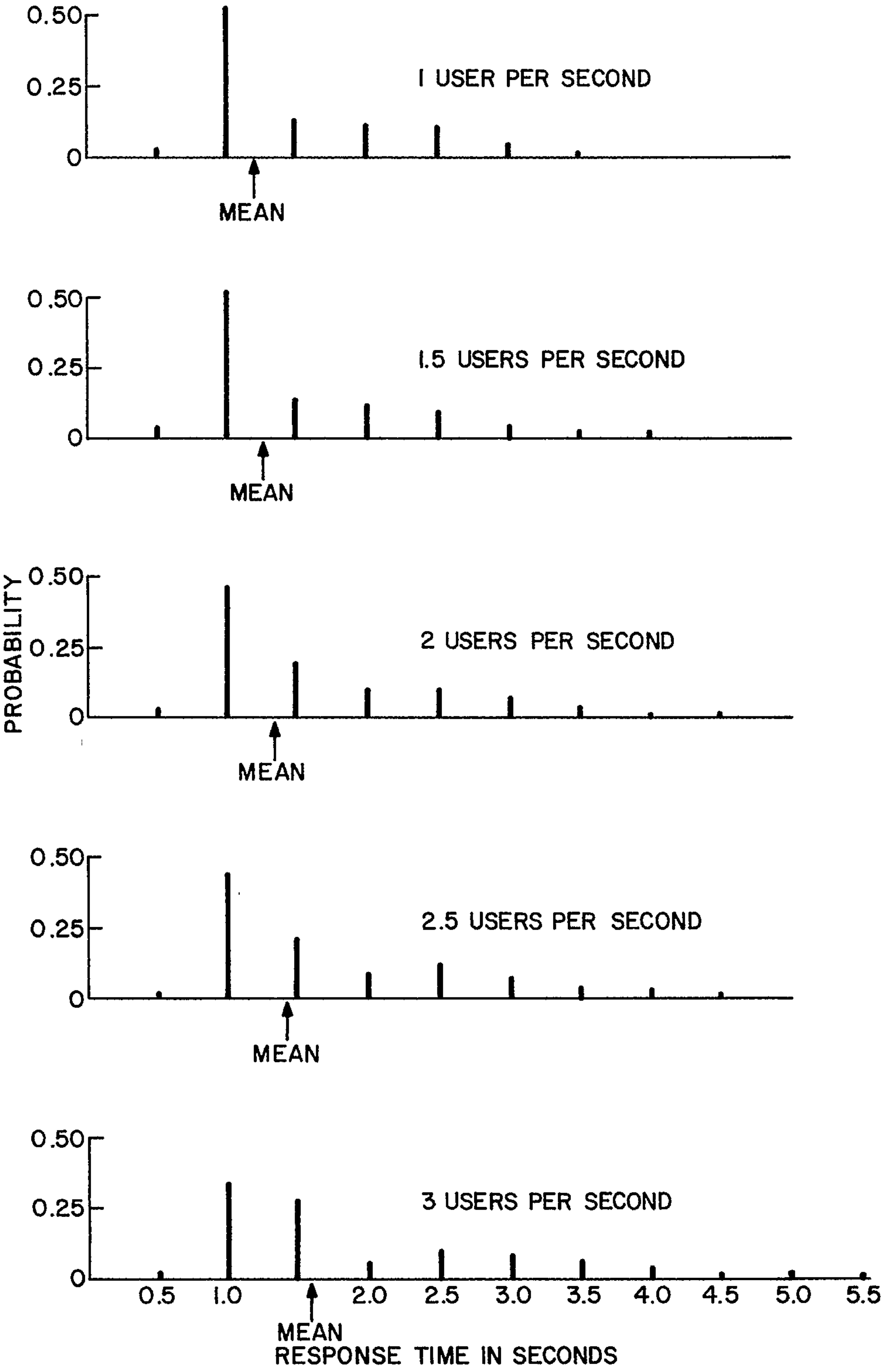


Exhibit 2A. Distribution of Response Times for Various Arrival Rates of Users (Case 1)

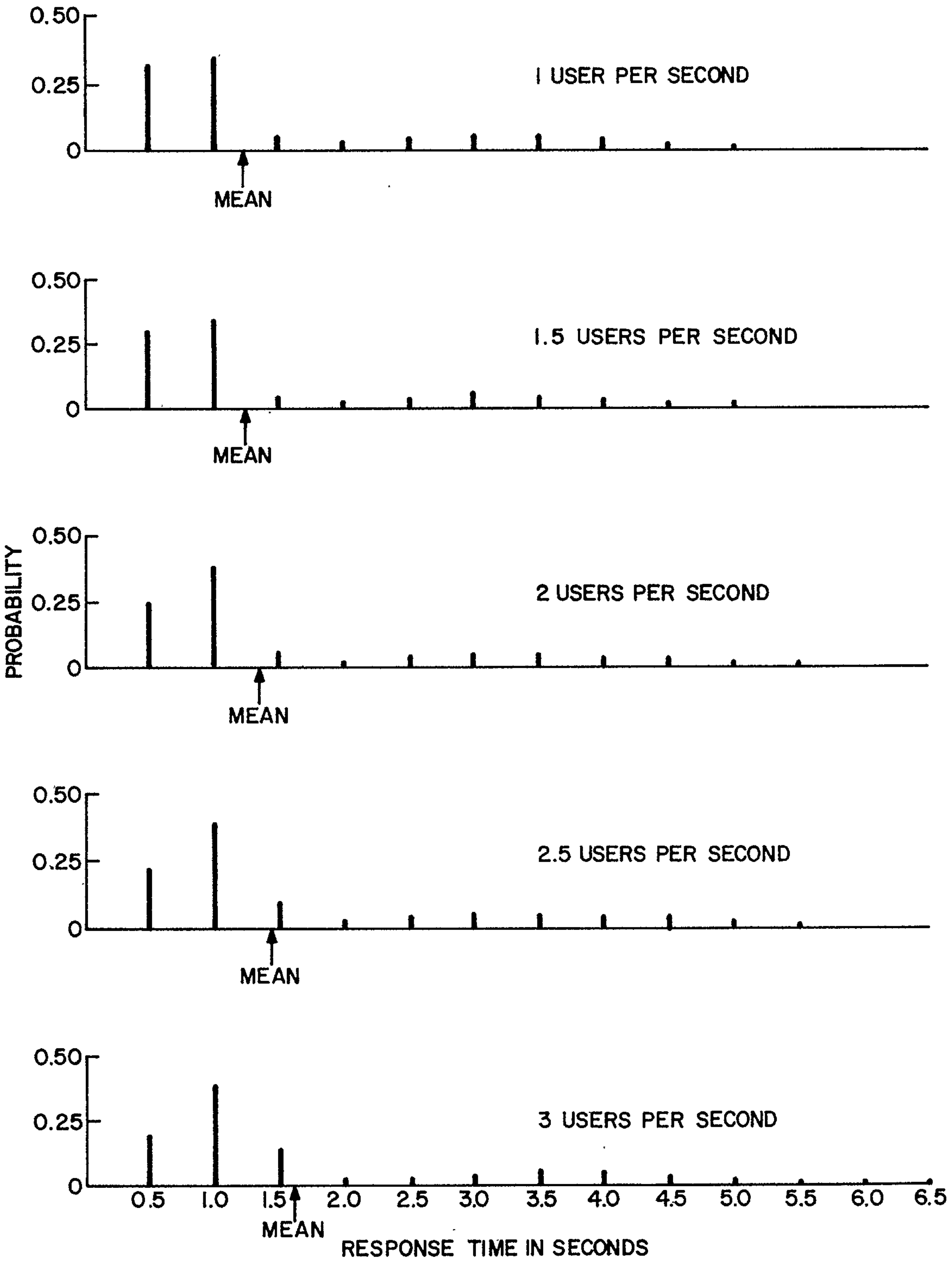


Exhibit 2B. Distribution of Response Times for Various Arrival Rates of Users (Case 2)

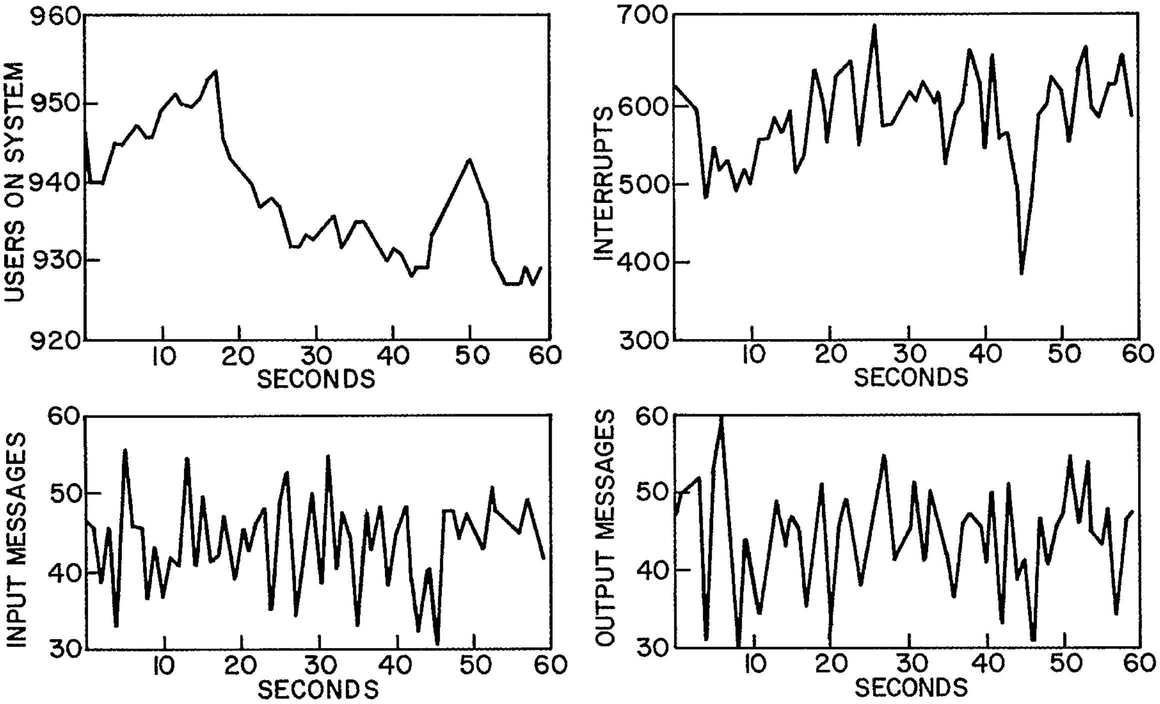


Exhibit 3. Dynamic Behavior of System For One Minute (Case 1)

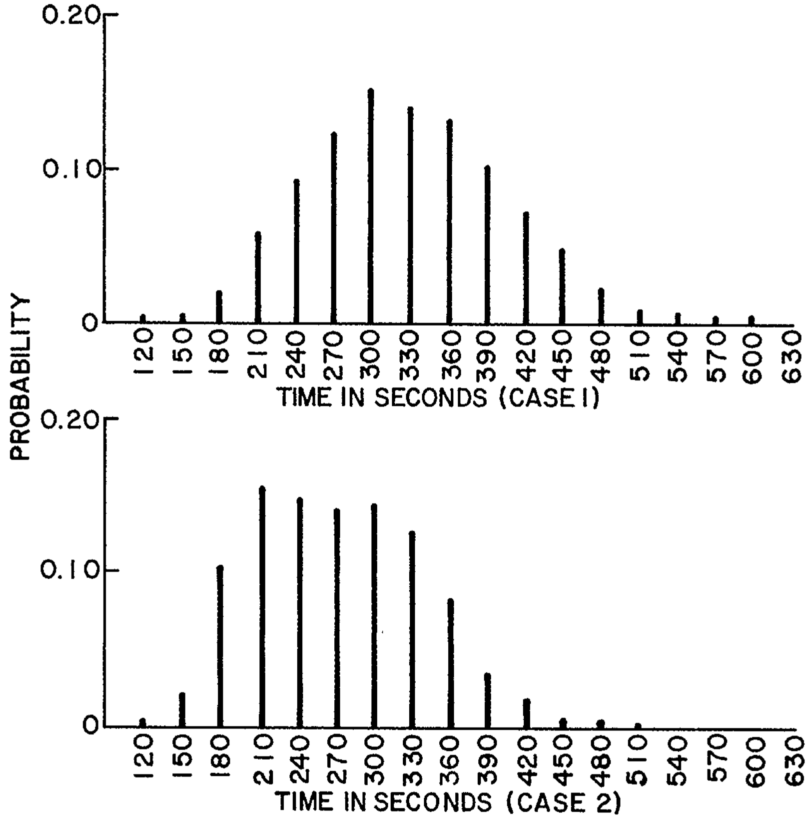


Exhibit 4. Distribution of User Holding Times

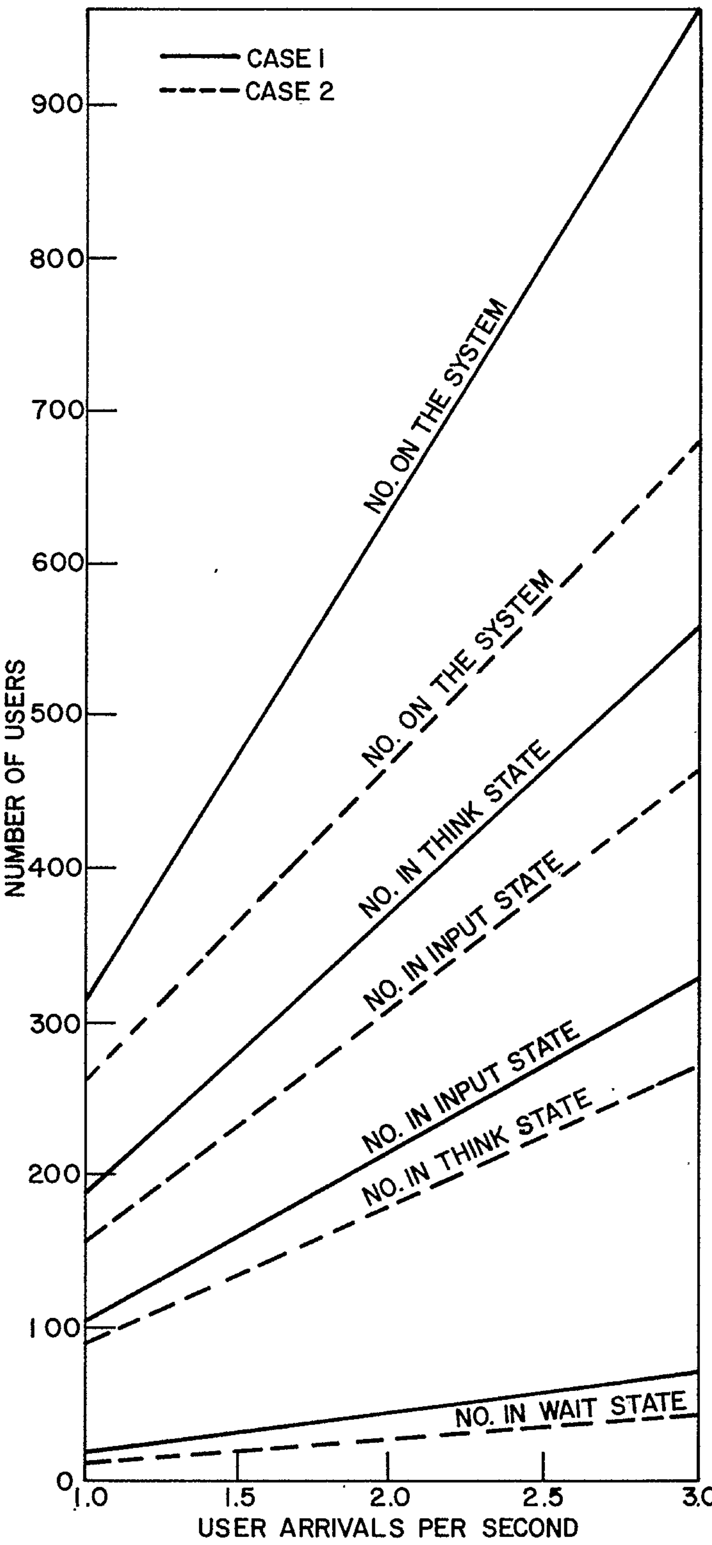


Exhibit 5A. User-State Distributions

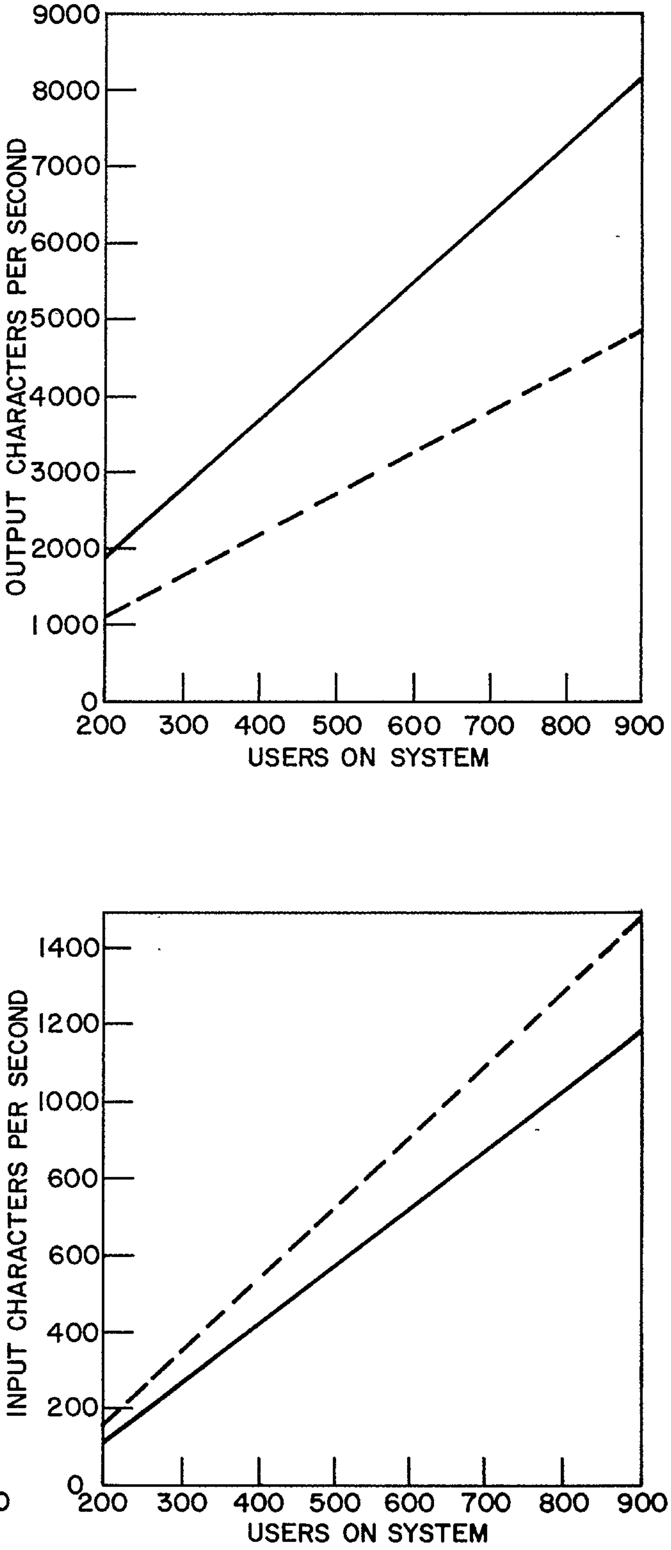


Exhibit 5B. Input-Output Rates

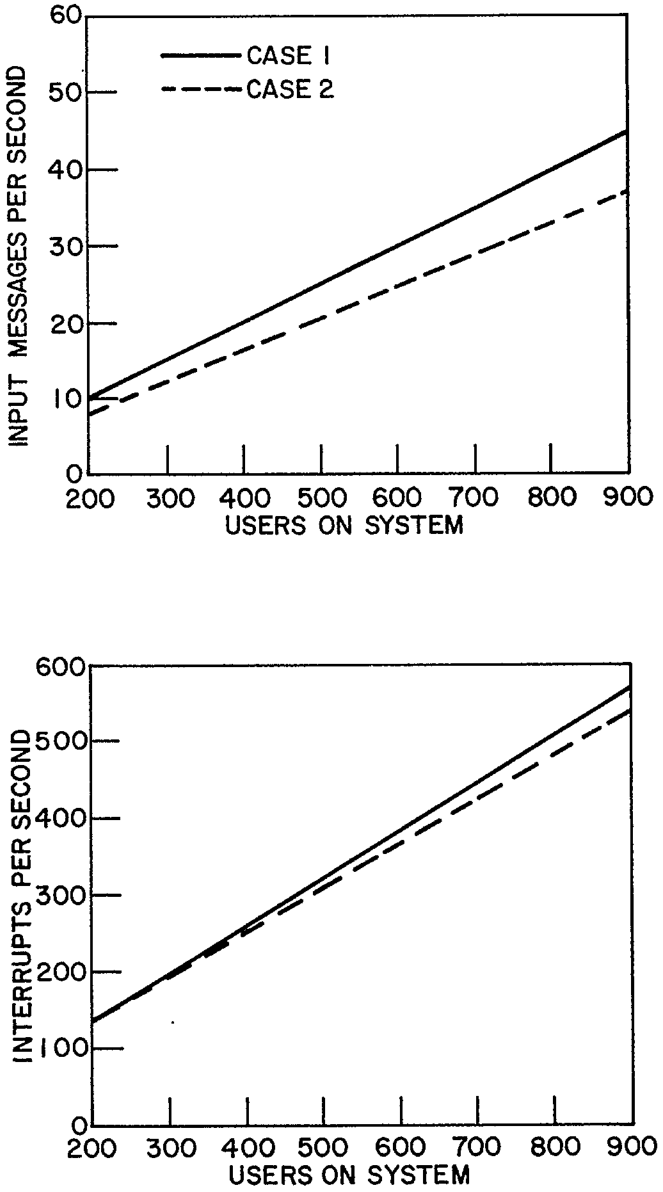


Exhibit 5C. Rates of Input Messages
and Interrupts

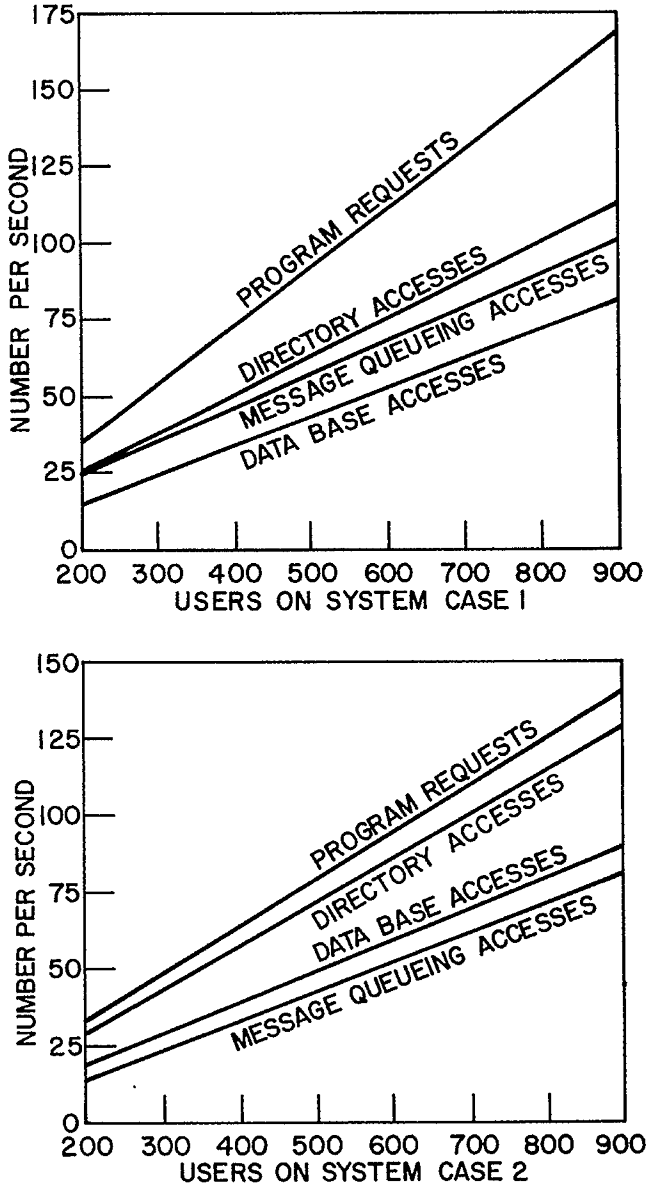


Exhibit 5D. Processing Rates

BIOGRAPHY

Harold Anderson is presently at Syracuse University, working toward a Ph. D. in Systems and Information Science. He is on educational leave from the IBM Thomas J. Watson Research Center, Yorktown Heights, New York, where he was a Research Staff Member involved in the development of systems analysis techniques for modeling interactive computer systems. Before he joined IBM in 1965, he worked for the United Aircraft Research Laboratories (1961), and Booz, Allen Applied Research Corporation (1964) as a programmer analyst, working on scientific and commercial applications of computers. Mr. Anderson has a B. S. in Engineering Mathematics (University of Rhode Island–1961), and M. S. in Physics (Trinity College–1965). He was born in New York City, August 11, 1939.