Performance of the Supermarket System is measured by throughput of the shoppers and the response time of the system to messages generated during checkout. This paper discusses some system design features adopted for the purpose of meeting a performance objective and two models developed for analyzing the throughput capacity of the system.

Store performance studies for the Supermarket System

by W. C. Metz, Jr. and D. Savir

The functional objective of the Supermarket System is to provide complete shopper checkout and store support facilities and to capture vital operational data at the point-of-sale. The system consists of a store controller and point-of-sale terminals that communicate with the controller through two store loops (maximum of 12 terminals per loop). The controller is capable of batch communication with a System/370 host computer on a switched line.[1] The controller can also provide back-up for a companion (remote) store whose controller has failed.[2] Communication with the terminals in the remote store is through the store loops in the remote store and a switched line[3] (for a detailed description of the system, see the paper by McEnroe, Huth, Moore, and Morris;[4] and of the scanner, see Dickson and Soderstrom[5]).

The performance of the system can be evaluated from two measurable variables. The first is *throughput,* the number of items or shoppers checked out per unit time; the second is *response time,* the time required for the system to respond to messages generated in the checkout operation. These two performance variables are not independent. Typically, as throughput increases, so does response time, and vice versa. However, a point of system overload will be reached if one or more of the various system queues (processor, disk, loop, etc.) becomes so congested that the time spent waiting in these queues (included in the response time) causes the system to delay the entry of subsequent messages. If response times are increased beyond the point of overload, the throughput will not reach the level that the system is

being asked to maintain. This situation is more likely to occur when the system is operating in back-up mode than when operating normally because back-up mode stresses the system the most (the controller performs the checkout functions for two stores[2]). Even though back-up mode is seldom used, the performance evaluation of this mode is of primary importance.

A performance objective of the system is to maximize the throughput capacity of the store subject to certain constraints, which necessitates keeping the response time below the point of system overload. In other words, the system should be *checker-paced*, enabling the checker (the person checking out a shopper's purchases) to scan and key information from items and other messages as fast as he wishes without being affected by responses from the system.

This paper presents an analysis of some system design features that were chosen to meet this performance objective and describes two tools developed for analyzing the system throughput capacity. The first tool is a GPSS model that simulates the principal elements of checkout in a peak load situation when the shopper queues are never empty. It is used to evaluate system design changes and to estimate, in back-up mode, the point of system overload (if it exists) and the throughput capacity when operating above the point of system overload. The second tool is an analytic model that, on the assumption that the system is operating below the point of system overload, solves the equilibrium queuing problem of determining the number of operational terminals required to obtain a specified level of throughput, subject to specific constraints on shoppers' waiting time. After taking the approximations of the models and the imprecision of data into account, the results from these models may help in assessing the ability of a particular system configuration to satisfy the requirements of a supermarket.

Performance factors

A more specific definition of response time is that interval initiated by the completion of data entry into the terminal, causing a message to be generated, and terminated by the completion of printing of the appropriate controller-generated message (response message) on the shopper's receipt tape. The response time depends upon specific constraints that are parameterized, so that the maximal system throughput is a function of the parameters chosen (which have some cost associated with them) enabling both the system designer and the user to evaluate and choose rationally from the available options. A set of these parameters, called a *store profile*, is determined partially by the human factors[6] of the checker and partially by the store policy and

the store neighborhood, its geography, its ethnic background, its average income, etc. Most of the parameters affect either the time to check an item or the time to process a shopper and thus determine the rate of message entry and the message mix.

The time to check an item depends upon:

- The checking device used—scanner or keyboard.
- The unloading discipline—whether the shopper or checker unloads.
- The type of checkstand used.
- The disposal of the item—whether the checker immediately bags it or whether he sends it elsewhere for immediate or subsequent bagging.

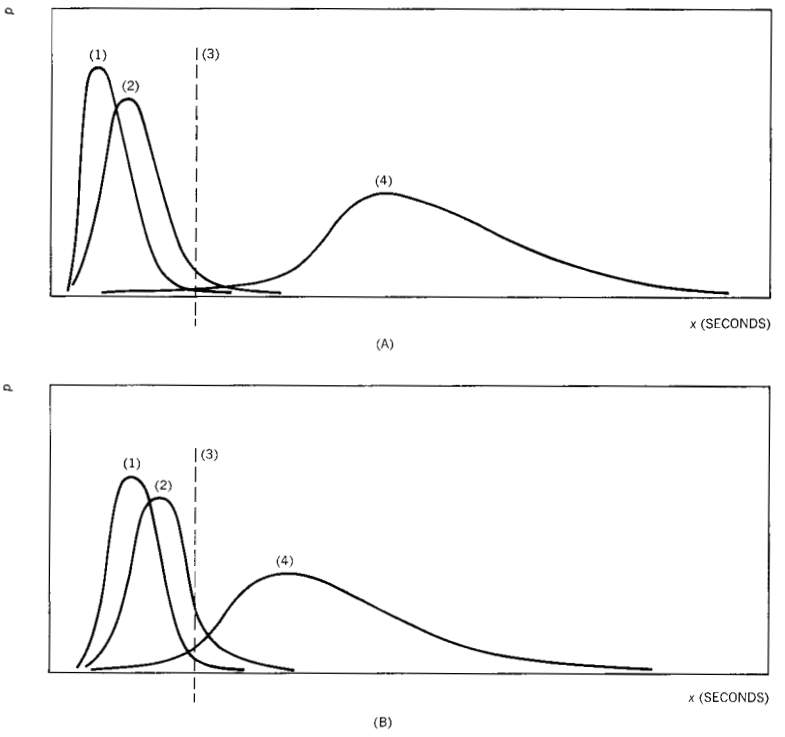
The time to process a shopper depends upon:

- The bagging discipline—whether or not the checker sets up and loads bags, and whether or not he assists in bagging.
- The method of handling checks—whether or not check authorization is required, and if so, whether or not it is done at the checkstand or at a service booth.
- The amount of personal attention given the shopper.
- The presence or absence of check tender, cash tender, food stamps tender, coupon tender, discounts, voids, and refunds.
- The equipment used—stamp dispenser, coin dispenser, scales, etc.

The performance objective of a checker-paced system imposes some specific response-time requirements. Total and tender entries require a rapid response because the checker cannot proceed with the transaction until the response message has been displayed. Item scan and item key entries require a rapid negative response so that the checker will not have disposed of the item if the system detects an error for that item. For this reason, as many errors as possible should be detected almost instantaneously at the terminal (e.g., scanning errors, invalid keying sequences, etc.). For the remaining errors that are detected at the controller (e.g., item not on file), the negative response should be rapid enough for the checker to easily retrieve and re-enter the item.

For each correct item-sale entry, a file look-up procedure must be executed, using the item code to retrieve the item price and description for printing and displaying at the terminal. It is not necessary in a checker-paced system that the positive response to each item-sale precede the scanning or keying of the next item-sale message from the same terminal. The Supermarket System is designed so that the checker-paced condition can normally be met by a weaker requirement for the distribution of

Figure 1 Overlap of printing, scanning, and loop/controller response time for a system in a normal (nonback-up) configuration, (A) Example of single-man mode of operation, (B) Example of double-man mode



LEGEND: (1) IS THE DENSITY FUNCTION FOR THE LOOP/CONTROLLER RESPONSE TIME FOR A SYSTEM OPERATING WITH 12 TERMINALS; (2) IS THE DENSITY FUNCTION FOR THE LOOP/CONTROLLER RESPONSE TIME FOR A SYSTEM OPERATING WITH 24 TERMINALS; (3) IS THE PRINT TIME; AND (4) IS THE DENSITY FUNCTION FOR THE SCANNING INTERARRIVAL TIME.

response times, by allowing the system to accept multiple (up to three) item-sale entries from the same terminal before completing the response to the first entry. This procedure is accomplished by providing each terminal with two transmit buffers, one receive buffer, and one print buffer. If three items are quickly scanned, the first item response message can be printing while the second is waiting to be received from the controller, and the third item message is in one of the transmit buffers waiting to be sent to the controller. Thus the printing portion of the response time to one message overlaps the succeeding message's portion attributable to the loop and controller. By partially overlapping the responses to successive messages, this double transmit-buffer scheme permits the checker to scan a few items at a rate much faster than his average scan rate.

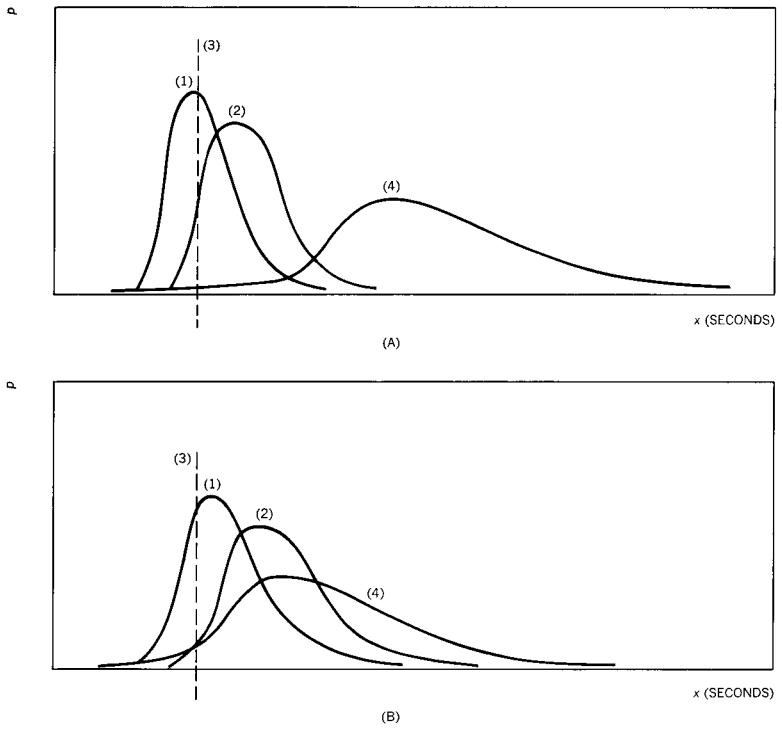
In Figure 1 are two graphs showing examples of the overlap of printing, scanning, and loop/controller response time. On the same axes are plotted the print time, two probability density functions for loop/controller response time, and a probability density function for scanning interarrival time (the time between successive scans). In Figure 1A, the density function for scan-

ning interarrival time is characteristic of a "single-man" mode of operation, that is, one person scanning and bagging (over-the-end checkstand).[6] In Figure 1B, the density function for scanning interarrival time is characteristic of a "double-man" mode of operation, that is, one person scanning while another is bagging.[7] Notice that the scanning interarrival times in the double-man mode are stochastically less than those in the single-man mode. As a result, if everything else were equal (the same timings for other checkout functions, the same distribution of items per order, the same number of terminals, the same terminal utilization, etc.), the double-man mode would impose a greater load on the system than the single-man mode, and the loop/controller response times in the double-man mode will be stochastically greater than in the single-man mode. Other curves may be characteristic of differing store profiles and system parameters; continuous variation of parameters generates continuous families of curves. For a checker-paced system *both* the print time *and* the loop/controller response times should be less than most of the scanning interarrival times. Notice that fewer response times exceed the interarrival times in the single-man mode (Figure 1A) than in the double-man mode (Figure 1B); but, in both cases, the storage provided by the double transmit-buffer scheme in the terminal accommodates short sequences of short interarrival times.

In back-up operation, the additional terminals on the controller will cause increased controller disk and controller microcode utilization, which will increase response times in both stores. Response times in the remote store will be increased to a greater extent than those in the local store because of (1) the greater contention for the line in the remote store resulting from the serializing of the two loops, making up to 24 terminals compete for the line in the remote store, compared to, at most, only 12 terminals competing for a line in the local store and (2) turnaround delay on the switched line from send to receive and vice versa. Because of batching both output and input messages to and from many terminals on a loop (induced by a single polling sequence[4]), the delay occurs not on each message but only on each batch.

Figure 2 shows examples of the increased response times in the remote store and thus the increased overlapping of printing, scanning, and loop/controller response time. Notice that, especially in the double-man mode, many of the response times are greater than the interarrival times. Normally the checker will not be affected by these longer responses because of the double transmit-buffer scheme. However, it is possible that the response to a particular message will be so long that the system will reject the entry of a subsequent message from the same terminal because there is no available transmit buffer. A "buffer overflow" error condition is created which the checker must "clear" before

Figure 2 Overlap of printing, scanning, and loop/controller response time for the remote store in a back-up configuration, (A) Example of single-man mode of operation, (B) Example of double-man mode



LEGEND: (1) IS THE DENSITY FUNCTION FOR THE LOOP/CONTROLLER RESPONSE TIME FOR 8 TERMINALS IN THE REMOTE STORE; (2) IS THE DENSITY FUNCTION FOR THE LOOP/CONTROLLER RESPONSE TIME FOR 16 TERMINALS IN THE REMOTE STORE; (3) IS THE PRINT TIME; (4) IS THE DENSITY FUNCTION FOR THE SCANNING INTERARRIVAL TIME.

again attempting to enter the message. When this occurs, the checker-paced condition is not met since the system is reducing the rate at which the checker is checking out groceries.

GPSS supermarket simulation model

The GPSS model has two primary uses: (1) to evaluate the throughput capacity of the system for all combinations of terminals from zero to 24 in the local store and zero to 24 in the remote store given a store profile, and (2) to evaluate design changes and functional enhancements in terms of their effect upon system throughput capacity.

The model has approximately 1400 GPSS blocks. The unit of simulated time is a fraction (1/8) of the bit time on the store loop so that the timing on the store loops can be simulated by whole units of simulated time. The ratio of simulated time to execution time on a System/360, Model 75 was observed to be in the range between one to four and one, depending upon the number of terminals and the store profile. Simulated run time is

normally in the 10 to 20 minute range, depending upon the profile parameters, particularly the variability of the number of items per order. The major components of the model are the checker, the terminal, the controller microcode, and the controller disk file.

checkers

Because of the checker-paced condition, the checkers must necessarily be a major component of the model. All the normal operations that the checker performs are simulated. These include the scanning of items with the UPC symbol; the tendering of cash, checks, food stamps, and coupons; change making; time between orders; the setting up and setting aside of bags; and the "clearing" and rescanning or rekeying when a "buffer overflow" condition occurs.

terminals

The terminal buffer scheme and print time are simulated. The terminal processing time, which is a small fraction of the response time, is approximated by a constant value.

store loops

Timings on the store loops are deterministic. The model records the utilization of the loop in terms of five percentages: (1) the time that the controller is transmitting the polling sequence,[4] (2) the time the controller is transmitting output messages, (3) the time the controller is receiving input messages. (4) the time the controller is performing none of the above due to a one-bit time delay incurred as the message stream is processed through a terminal, and (5) the time the controller is doing none of the above due to the turnaround time or the propagation delay in the back-up transmission (propagation delay includes modem delay in the two stores and the delay of the switched line). The turnaround time and the propagation delay can be varied to see their effect on throughput capacity in the remote store.

controller microcode

The controller microcode is modeled by counting microinstructions and assuming an average execution time per microinstruction. All queues are simulated and the logic in the model of servicing these queues is similar to the logic in the controller. Interruptions are simulated with minor approximations so that controller utilization by interruption level is measured.

controller disk

The arm movement of the disk is simulated from the disk seek function (seek time as a function of the number of cylinders traversed) by randomly choosing a cylinder in the appropriate file. Latency is simulated by randomly choosing a sector when the sector location is not known, or by calculating the latency time when the sector location is known. The latter situation occurs in the write-back sequence of a file update operation. For the price-description and check-authorization random files, the chaining due to synonyms is simulated. The probability of chaining in the price-description file is a function of the UPC assign-

ments,[8] the file size, the percent loading of the file, the method of file loading, and the randomizing algorithm. The percent loading is a variable in the model.

input

In addition to simulating fixed or deterministic time intervals and the delays due to waiting for hardware and software resources, some random events are also simulated. For these events, however, there are inherent errors in simulation due to imperfect knowledge of their probability distributions. All the checker operations in the model are probabilistic events, and probability distribution functions are required to simulate them. For some of the operations, specifically scanning and keying, frequency distributions of interarrival times are available from human performance testing.[6] These distributions all have the same general shape. Each is unimodal and positively skewed and can be approximated by a gamma distribution with an offset. Each distribution then requires three parameters (including the offset) which determine its shape. However, each gamma distribution may be transformed into a normalized gamma distribution with only one parameter, called the alpha parameter (see Appendix A). The advantage in this is that random variables from many different distributions required in the model can be obtained from the same normalized distribution by generating a random variable from the normalized distribution, and then transforming it into a random variable from the original distribution (see Appendix B). Thus, most of the probabilistic time intervals in the model are generated from one of five normalized gamma distributions, each with a different alpha parameter.

The input to the model consists of the parameters of a particular store profile. Some of the parameters affect the message mix—for example, those specifying the number of items per order, the proportion of items that are multipriced, the proportion of orders with coupons, the proportion of orders with food stamps, and the proportion of orders with check authorization. Other parameters affect the message input rate by specifying the interarrival time distribution of scanned and keyed messages, the distribution of change-making time, and the distribution of bag-setup and bag-aside time. For each time interval, a minimum, a mean, and an alpha parameter, characterizing a particular gamma distribution, are required. The mean of these time distributions determines an inherent average throughput rate for the store profile, which can be expressed in messages per second per terminal. The average throughput rate best describes the load on the system, because it indicates store loop utilization and, to some extent, disk utilization. Other parameters that influence disk utilization provide additional information on the amount of system load inherent in a given profile. Examples are frequency of multiprice items and check authorization, both of which require considerable disk utilization.

output

The output includes standard simulation statistics such as the utilization of hardware and microcode resources and distributions of waiting times, service times, and response times. These statistics are useful for analyzing the system response time for a given profile and number of terminals, and for evaluating design changes. The output also includes empirical distributions of the probabilistic time intervals to check convergence to the distributions predicted by the input parameters of the gamma distribution functions.

throughput curves

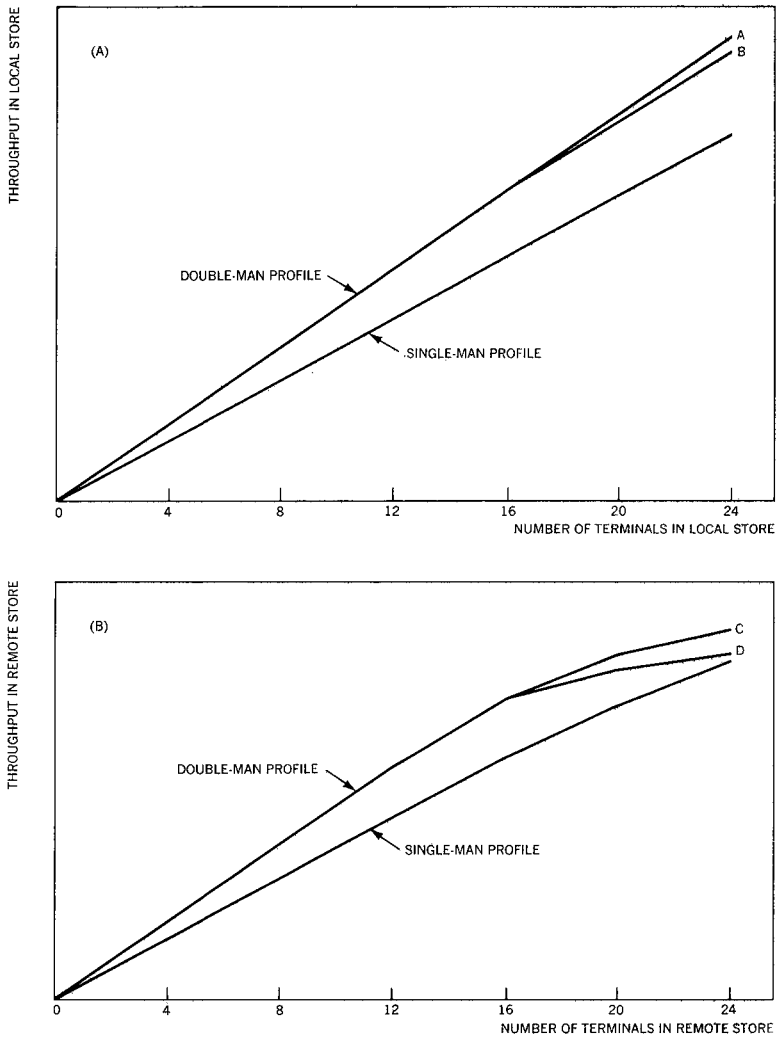
For model runs of the back-up mode, an important result is the throughput capacity of the local and remote stores. With just a few model runs, the throughput in the local and remote stores can be plotted as a function of the number of terminals in the local and remote stores, respectively. Notice the curves labeled "single-man profile" in Figures 3A and 3B. For this particular profile, the system load is low enough so that the throughput in the remote store is independent of the number of terminals in the local store, and vice versa. The slope of the curves labeled "single-man profile" in Figure 3A and in Figure 3B, in the range between zero and 16 terminals, is a constant equal to the inherent throughput rate per terminal of the store profile. Notice in Figure 3B that the slope of the curve of the remote store decreases slightly beyond 16 terminals although the total throughput continues to rise.

Similar graphs obtained from model runs of a particular double-man profile are also shown in Figures 3A and 3B. The slopes of these curves are greater than those for the single-man profile because the inherent throughput rate of the double-man profile is greater than that of the single-man. Also, the system load is high enough, with many terminals operating in both stores, that the throughput capacity in each store is a function of the number of terminals in both stores, hence, a function of two variables. For simplicity, we indicate only two functions of the second variable, i.e., the number of terminals in the remote store when the local store throughput capacity is shown and vice versa. In Figure 3A, the curves labeled A and B are for 12 and 24 terminals, respectively, in the remote store. In Figure 3B, the curves labeled C and D are for 12 and 24 terminals, respectively, in the local store.

Supermarket queuing model

When installing a supermarket system, the customer must trade off decisions as to his checkout procedures (e.g., scanners or key entry; over-the-end or double belt; baggers or none; check processing on- or off-line; etc.) against the number of checkstands that he will need to provide a specified level of service to the shopper. This model permits the estimation of the number of checkstands required as a function of the parameters that are

Figure 3 Throughput capacity



determined by the checkout procedures and by the profile of shoppers at the store. These parameters are (a) time to check an item, (b) time to process a shopper, and (c) distribution of items per shopper.

Shopper service level is defined in terms of conditional waiting time, or the waiting time of those shoppers who must wait in line at the checkstands for service. Two levels are set—a mean level, C_M, specifying a maximum average conditional waiting time, and a 90-percentile level, C_{90}, specifying that waiting time that no more than 10 percent of the shoppers who wait may exceed.

The maximum capacity of the store in terms of shoppers per hour is then calculated for a variable number of checkstands as a

function of the parameters and the constraints on service level. The evaluation of the effect of choice of parameters on the number of checkstands required to handle the estimated load can be made.

The service process is a composite of two processes—a random number of items are checked and the shopper himself is processed afterward. The distribution of service time will therefore depend on the distributions of time to check an item, time to process a shopper, and the number of items associated with that shopper. These variables are mutually independent.

The time to check an item is the elasped time between the instant at which the checker first grasps an item and the instant at which he first grasps the next item. If produce is weighed at the checkstand, or if any other item-related activity interrupts the general checking process, it is best to count the time spent in such interruptions as part of the time spent processing the shopper.

The time to complete the processing of a shopper before and after the items are checked may include the greeting of the shopper, moving of items on the checkstand belt, bag setup, bagging after checking is finished (either alone or with a bagger), loading of full bags, time spent in interruptions to the item checking, presenting the bill, receiving payment and giving change, waiting for the shopper to write a check, processing the check, etc. The mean time to process an item is denoted by μ^{-1}. The mean time to process a shopper is denoted by σ^{-1}. μ and σ are taken to be the rates of respective negative exponential distributions. (Negative exponential distributions were chosen for mathematical tractability; a three-parameter gamma distribution is more representative of scanning (Figures 1 and 2), whereas completely different distributions characterize other means of data entry.)

The distribution of items per shopper is taken to be the negative binomial with mean m and parameter p. A discussion of the applicability of this distribution and the estimation of its parameters from store data is given in Reference 9. The negative binomial distribution is either (a) unimodal with left and right tails or (b) decreasing to the right, depending on the values of its parameters. The distributions of items per shopper in supermarkets have been seen to exhibit both these characteristics, depending on a variety of factors such as local buying habits, parking facilities, peak hour average order size, proximity of convenience stores, time of day, etc. In general, the likelihood of the presence of many large orders will tend to deter shoppers with small orders (producing a distribution of type a), but most shopping trips are for small orders (producing a distribution of type b).

The service distribution of the shopper at the checkstand is accordingly described in Reference 9. The queuing model of the supermarket is represented by an $M/G/k$ queue with Poisson arrivals, and k servers or checkstands, each of which has this service distribution. The model is applicable in an environment in which the linear portion of the throughput curves (Figure 3) applies.

The $M/G/k$ queue does not exactly describe the front end of a supermarket; if it did, then each arriving shopper would have the prescience to select that checkstand which would minimize his waiting time. However, on the whole, most shoppers make a good decision, especially since the opportunity for jockeying exists, at least until the merchandise is unloaded on the checkstand. The supermarket almost has one property of the $M/G/k$ queue: no checkstand will have shoppers waiting while another checker is idle (unless an unfortunate waiting shopper has already begun to unload his items on the checkstand; his waiting time will tend to be longer than that predicted by the model, but he will reproach himself for his own poor judgment rather than fault the store's service level).

The $M/G/k$ queue is not readily given to analysis. The behavior of the $M/G/1$ queue is well known, but the representation of k checkstands at the front end of a store by k distinct $M/G/1$ queues is very unsatisfactory, since it would imply that each shopper picks a checkstand at random regardless of the queue lengths at all the checkstands. However, there is a result which leads to a useful approximation:[10] in an $M/M/k$ queue (Poisson arrivals, exponentially distributed service time, k servers), the distribution of waiting time of those who wait is exactly the same as the distribution of waiting time of those who wait in an $M/M/1$ queue whose server works k times as fast. Of course, the number who must wait is not the same.

The subject of interest in our $M/G/k$ queue is precisely the waiting time of shoppers who must wait, for this determines and is determined by the store's service level. Hence, with respect only to the distribution of waiting time of those shoppers who must wait (conditional waiting time), the $M/G/k$ queue can be approximated by an $M/G/1$ queue whose server works k times as fast.

The capacity of a store with mean, C_M, and 90-percentile, C_{90}, constraints on conditional waiting time is the maximum arrival rate of shoppers per checkstand, λ, that will satisfy both

$$\text{mean conditional waiting time} \leq C_M$$

and

$$\text{90-percentile conditional waiting time} \leq C_{90} \qquad (1)$$

Figure 4 Program input and output of example

```
      STORECAPACITY
ENTER LEAST NUMBER OF CHECKSTANDS
?:
      1
ENTER GREATEST NUMBER OF CHECKSTANDS
?:
      15
ENTER MEAN NUMBER OF SECONDS TO CHECK ITEM
?:
      2
ENTER MEAN NUMBER OF SECONDS TO PROCESS SHOPPER
?:
      30
ENTER MEAN NUMBER OF ITEMS/SHOPPER
?:
      13.56
ENTER VALUE OF 'P'
?:
      .24
ENTER MEAN CONSTRAINT IN MINUTES
?:
      7
ENTER 90-CENTILE CONSTRAINT IN MINUTES
?:
      7

     1          49.87
     2          112.8
     3          175.8
     4          238.8
     5          301.8
     6          364.8
     7          427.8
     8          490.9
     9          553.9
    10          616.9
    11          679.9
    12          742.9
    13          806
    14          869
    15          932
```

An adequate approximation to the 90-percentile is given by the mean plus 1.3 standard deviations.

In Reference 9 we show that constraints in Expression 1 are equivalent to

$$\lambda \leq \frac{kC_M - A}{kC_M b_1} \tag{2}$$

$$\lambda \leq \frac{1.69 + 2kC_{90}(kC_{90} - A)b_1 - 1.3[0.69B^2 + (B - 2kAC_{90}b_1)^2]^{\frac{1}{2}}}{2k^2 C_{90}{}^2 b_1{}^2}$$

where

$$A = \frac{b_2}{2b_1}$$

and

$$B = \frac{b_2^2}{2b_1} - \frac{b_3}{3}$$

and b_j is the jth service time moment:

$$b_1 = \frac{m}{\mu} + \frac{1}{\sigma}$$

$$b_2 = \frac{m(1+p)-1}{p\mu^2} + \frac{1}{\sigma^2} + \left(\frac{m}{\mu} + \frac{1}{\sigma}\right)^2$$

$$b_3 = 2\left[\frac{m(1+p+p^2)-(1+p)}{p^2\mu^3} + \frac{1}{\sigma^3}\right] + 3\left(\frac{m}{\mu} + \frac{1}{\sigma}\right)\left(\frac{m(1+p)-1}{p\mu^2} + \frac{1}{\sigma^2}\right) + \left(\frac{m}{\mu} + \frac{1}{\sigma}\right)^3$$

An APL/360 program was written to perform the above calculations.

The conclusion of this section is an example illustrating the use of the methodology. Consider a store with the following profile: the mean number of items per shopper is 13.56 and the variance is 51.28. Then $p = 12.56/51.28 = 0.24$.[9] Suppose that the mean time to check an item is two seconds and the mean time to complete the processing of a shopper is 30 seconds. The number of checkstands required to handle a volume of 400 shoppers per hour subject to a 90-percentile constraint of seven minutes waiting time for service is sought. Figure 4 illustrates the input and output of the APL program.

Clearly, seven checkstands will suffice. Now suppose the store wished to use an express lane with a limit of 10 items per shopper. The model is used to investigate the effects of this change.

Assume that the mean number of items per shopper in the express lane is five with a variance of 15. Suppose that the mean time to check an item is three seconds, since a slower checkout device might be used on an express lane; and suppose that the mean time to complete the processing of a shopper is 20 seconds, since there will be less check cashing and bag handling in the express lane. The express lane capacity is therefore 103 shoppers per hour. $400 - 103 = 297$ shoppers will be handled by the remaining regular checkstands. The express lane is assumed

Figure 5 Program input and output when express lane is considered

```
      STORECAPACITY
ENTER LEAST NUMBER OF CHECKSTANDS
□:
      1
ENTER GREATEST NUMBER OF CHECKSTANDS
□:
      7
ENTER MEAN NUMBER OF SECONDS TO CHECK ITEM
□:
      2
ENTER MEAN NUMBER OF SECONDS TO PROCESS SHOPPER
□:
      33.47
ENTER MEAN NUMBER OF ITEMS/SHOPPER
□:
      16.53
ENTER VALUE OF 'P'
□:
      .52
ENTER MEAN CONSTRAINT IN MINUTES
□:
      7
ENTER 90-CENTILE CONSTRAINT IN MINUTES
□:
      7

     1          41.6
     2          95.58
     3         149.7
     4         203.7
     5         257.8
     6         311.9
     7         366
```

never to be empty—if the queue there is too long, shoppers will not join it, preferring to choose a regular checkstand; if there is nobody waiting in the queue, a shopper with more than 10 items will sneak in.

The mean number of items per shopper in the regular checkstands is

$$[(400)(13.56) - (103)(5)]/297 = 16.53$$

The variance is

$$\frac{(400)(51.28 + (13.56)^2) - (103)(15 + (5)^2)}{297} - (16.53)^2 = 29.59$$

The value of p consistent with this new mean and variance is $15.53/29.59 = 0.52$. The mean time to complete the processing of a shopper is

$$[(400)(30) - (103)(20)]/297 = 33.47 \text{ seconds}$$

Using these parameters, we find the number of checkstands necessary to handle a volume of 297 shoppers. The input and output of the APL program are shown in Figure 5.

Six checkstands will be necessary. The total number of checkstands has not been affected, although one regular checkstand has been replaced by a cheaper express checkstand. The overall service level is unchanged, although shopper satisfaction may be improved to the extent that the shoppers who wait tend to have more items under the express-lane system than otherwise, and, presumably, a shopper with more items will be more willing to wait than one with fewer items.

CITED REFERENCES AND FOOTNOTES

1. In the United States and Canada, this is a 2400 bits per second line. In other countries, local communications facilities dictate a 2400, 1200, or 600 bits per second line.
2. R. O. Hippert, L. R. Palounek, J. Provetero, and R. O. Skatrud, "Reliability, availability, and serviceability design considerations for the Supermarket and Retail Store Systems," in this issue.
3. In the United States and Canada, this is a switched line. In other countries, a leased line may be required.
4. P. V. McEnroe, H. T. Huth, E. A. Moore, and W. W. Morris, III, "Overview of the Supermarket System and the Retail Store System," in this issue.
5. L. D. Dickson and R. L. Soderstrom, *The IBM Supermarket Scanner*, Technical Report (in preparation), IBM Corporation, System Development Division, Rochester, Minnesota.
6. D. C. Antonelli, "The role of the operator in the Supermarket and Retail Store Systems," in this issue.
7. The use of the single-man and double-man mode is expected to be widespread in the United States and Canada. Other modes may, of course, exist and will be associated with the appropriate graphs.
8. D. Savir and G. J. Laurer, "The characteristics and decodability of the Universal Product Code symbol," in this issue.
9. D. Savir, *On the Number of Terminals Required for a Level of Customer Service in a Supermarket System*, Technical Report TR29.0124, IBM Corporation, System Development Division, Research Triangle Park, North Carolina.
10. D. R. Cox and W. L. Smith, *Queues*, John Wiley & Sons, Inc., New York, New York (1961).

Appendix A: Derivation of the normalized gamma distribution

The gamma density function is defined as

$$f(x) = \begin{cases} 0, & x < 0 \\ \dfrac{x^{\alpha-1}e^{-x/\beta}}{\beta^{\alpha}\Gamma(\alpha)}, & x \geq 0, \alpha > 0, \beta > 0 \end{cases}$$

where

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1}e^{-x}dx$$

The gamma function for random variable X is then

$$\text{Prob } (X \leq x) = F(x) = \begin{cases} 0, & x < 0 \\ \displaystyle\int_0^x \frac{t^{\alpha-1}e^{-t/\beta}}{\beta^{\alpha}\Gamma(\alpha)}\,dt, & x \geq 0,\ \alpha > 0,\ \beta > 0 \end{cases}$$

To derive a normalized gamma distribution function, define a second random variable Z by $Z = X/\beta$, with density function g and distribution function G. G in terms of F is

$$G(z) = \text{Prob } (Z \leq z) = \text{Prob } (X \leq \beta z) = F(\beta z)$$

g in terms of f is

$$g(z) = G'(z) = \beta F'(\beta z) = \beta f(\beta z),\ \beta z \geq 0$$

Thus the normalized distribution function G is

$$G(z) = \int_0^z g(t)dt = \int_0^z \beta f(\beta t)dt = \int_0^z \frac{t^{\alpha-1}e^{-t}}{\Gamma(\alpha)}\,dt$$

Notice that G is a function only of α and not of β. To derive a gamma distribution function with an offset as a function of G, define another random variable Y by $Y = \gamma + \beta Z$, where γ is the offset or minimum value of Y.

Then

$$\text{Prob } (Y \leq y) = \text{Prob } (Z \leq (y-\gamma)/\beta) = G((y-\gamma)/\beta)$$

Let R be uniformly distributed on [0, 1] and let G^{-1} be the inverse of G. Let the random variable $\hat{Z} = G^{-1}(R)$.

Then

$$\begin{aligned} \text{Prob } (\hat{Z} \leq z) &= \text{Prob } (G^{-1}(R) \leq z) \\ &= \text{Prob } (R \leq G(z)) \\ &= G(z) \end{aligned}$$

Therefore $\hat{Z}$ and Z are identically distributed and can be generated by the function G^{-1} on the random variable R.

G^{-1} can be put in the GPSS model. During model execution, GPSS will generate a random variable R which is any multiple of 0.001 from 0 to 0.999. R will be used to generate a random variable Z by G^{-1}, and then Z will be used to generate a random variable Y by

$$Y = \gamma + \beta Z \tag{A1}$$

The random variable Y obeys the probability law of a gamma distribution with a given α, β, and γ. Because G^{-1} is a function only of α, G^{-1} for a given α can be used to generate random variables from any gamma distribution with the same value for α, regardless of the values for γ and β.

Appendix B: Generating random variables from a normalized gamma distribution

The mean μ and mode ν of a gamma distribution with an offset are functions of the parameters α, β, and γ:

$$\mu = \gamma + \alpha\beta \tag{B1}$$

$$\nu = \gamma + \beta(\alpha - 1) \tag{B2}$$

Solving Equations B1 and B2 for α and β in terms of μ, ν, and γ gives:

$$\alpha = \frac{\mu - \gamma}{\mu - \nu}$$

$$\beta = \mu - \nu$$

Consider two frequency distributions for operator scanning interarrival times with parameters μ_1, ν_1, γ_1, and μ_2, ν_2, γ_2, respectively:

$\gamma_1 = 0.2$ second $\quad \gamma_2 = 0.2$ second

$\nu_1 = 1.5$ seconds $\quad \nu_2 = 1.1$ seconds

$\mu_1 = 2.2$ seconds $\quad \mu_2 = 1.57$ seconds

The two distributions have the same value for α:

$$\alpha_1 = \frac{\mu_1 - \gamma_1}{\mu_1 - \nu_1} = \frac{2.2 - 0.2}{2.2 - 1.5} = 2.9$$

$$\alpha_2 = \frac{\mu_2 - \gamma_2}{\mu_2 - \nu_2} = \frac{1.57 - 0.2}{1.57 - 1.1} = 2.9$$

Thus the same inverse distribution function $G^{-1}(\alpha = 2.9)$ may be used to generate random variables for the simulation of each of the respective scanning operations. In both cases, $G^{-1}(\alpha = 2.9)$ is used to generate a random variable $Z(\alpha = 2.9)$. The random variable Y_1 for the first scanning operation would be determined by substituting values for γ_1, μ_1, ν_1 into Equation A1:

$$\begin{aligned} Y_1 &= \gamma_1 + \beta_1 Z(\alpha = 2.9) \\ &= \gamma_2 + (\mu_2 - \nu_2) Z(\alpha = 2.9) \\ &= 0.2 + 0.7\ Z(\alpha = 2.9) \end{aligned}$$

The random variable Y_2 for the second scanning operation would be determined by substituting values for γ_2, μ_2, ν_2 into A1:

$$\begin{aligned} Y_2 &= \gamma_2 + \beta_2 Z(\alpha = 2.9) \\ &= \gamma_2 + (\mu_2 - \nu_2) Z(\alpha = 2.9) \\ &= 0.2 + 0.47\ Z(\alpha = 2.9) \end{aligned}$$

Reprint Form No. G321-5004