FTD-HC-23-1506-71

AD735641

FOREIGN TECHNOLOGY DIVISION

THIS IS AN UNEDITED ROUGH DRAFT TRANSLATION BY JOINT PUBLICATIONS RESEARCH SERVICES

APPLICATION OF THE QUEUEING THEORY TO THE INVESTIGATION OF INFORMATION SYSTEMS

by

M. Libura

Approved for public release;
Distribution unlimited.

19

UNCLASSIFIED

Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author) Foreign Technology Division Air Force Systems Command U. S. Air Force	2a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED 2b. GROUP

3. REPORT TITLE

APPLICATION OF THE QUEUEING THEORY TO THE INVESTIGATION OF INFORMATION SYSTEMS

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

Translation

5. AUTHOR(S) (First name, middle initial, last name)

Libura, M.

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
1970	15	18

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
b. PROJECT NO.	FTD-HC-23-1506-71
c. DIA Task Nos. T71-05-09 and d. T71-05-13	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report) AP1024332

10. DISTRIBUTION STATEMENT

Approved for public release; distribution unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
	Foreign Technology Division Wright-Patterson AFB, Ohio

13. ABSTRACT

This paper discusses the literature on application of the methods of queueing theory to the analysis of information systems. The principles of queueing theory are given, as are the most widely used results as applied to the simplest queueing systems. Priority systems are discussed and a simple example of their use is given. A survey of models of multi-access systems with time sharing is given formulated in the language of queueing theory.

DD FORM 1 NOV 65 1473

UNCLASSIFIED

Security Classification

KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Information Storage and Retrieval Data Processing Algorithm Computer Input						

UNEDITED ROUGH DRAFT TRANSLATION

by Joint Publications Research Services

APPLICATION OF THE QUEUEING THEORY TO THE INVESTIGATION OF INFORMATION SYSTEMS

By: M. Libura

English pages: 15

Source: Archiwum Automatyki i Telemechaniki (Records of Automation and Telemechanics) 1970, Vol. 15, No. 4, pp. 485-497

PO/0031-70-015-004

PREPARED BY:

TRANSLATION DIVISION
FOREIGN TECHNOLOGY DIVISION
WP-AFB, OHIO.

THE APPLICATION OF QUEUEING THEORY TO THE INVESTIGATION OF INFORMATION SYSTEMS

[Article by Marek Libura, Independent Information Processing Laboratory, Institute of Automation, Polish Academy of Sciences; Archiwum Automatyki i Telemechaniki, Vol 15, No 4, 1970, pp 485-497]

This paper discusses the literature on application of the methods of queueing theory to the analysis of information systems. The principles of queueing theory are given, as are the most widely used results as applied to the simplest queueing systems. Priority systems are discussed and a simple example of their use is given. A survey of models of multi-access systems with time sharing is given formulated in the language of queueing theory.

1. INTRODUCTION

Queueing theory is an apparatus which is particularly useful in analyzing the specific features of information systems. The complexity of these systems and the lack of complete information concerning their elements usually do not permit a deterministic description. In some cases a description in terms of queueing theory is the only possible one (as in the case of multi-access systems, where there is little information concerning users and where such information is usually available in the form of statistics concerning the rate of arrival of problems to the system).

There is at the same time a serious need for analytic results related to these systems. For example, in designing multi-access systems experience has shown that very detailed evaluations must be made of the effects of individual parameters on the operation of the system (such as the effect of the time quantum allocated on the throughput of the system, and the like). This type of evaluation is not always possible, or not always convenient, using modeling methods.

2. BASIC CONCEPTS IN QUEUEING THEORY

Queueing theory considers the model of an isolated queueing system as shown in Figure 1. At discrete moments of time $t_1, t_2, \ldots, t_n, \ldots$ the queueing

system receives problems making up the input stream. A complete description of the output stream requires giving the sets $\{\tau_i\}$, $i = 0, 1, 2, \ldots$, where $\tau_i = t_{i+1} - t_i$, $i\{\nu_i\}$, where ν_i denotes the number of problems arriving at the queueing system at moment t_i.

An equivalent description of the input stream consists in giving the function x(t) whose value at moment t is the number of problems which arrived at the queueing system during time [0, t]. Research is usually limited to considering stationary streams, i.e., streams in which the probability of the arrival of k problems in time segment [T, T + t] does not depend on T and is only a function of k and t.

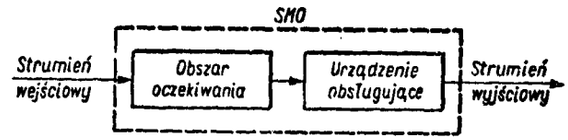

Figure 1. Isolated queueing system.

Strumien wejsciowy = input stream
Obszar oczekiwania = waiting area
Urzadzenie obslugujace = service equipment
Strumien wyjsciowy = output stream
SMO = queueing system

Problems arriving at the queueing system proceed to the waiting area. When the service equipment is free service begins immediately. If not, a problem queue is formed. The selection of problems to service is determined by the so-called waiting discipline.

Among the applications considered two types of service discipline are interesting. These are service in the order of arrival (FIFO, or first-in-first-out, FCFS, or first-come-first-service), and service in accordance with a priority system. Most of the queueing systems considered here are priority systems.

The basic value which characterizes the service equipment is the distribution of service time for each problem. Other parameters are the number of service channels (designated hereinunder with the symbol m) and the number of service steps.

When information systems are examined individual elements in the queueing model assume different interpretations. The input stream usually corresponds to incoming programs and individual instructions. The waiting area corresponds to buffer memory, operating memory, and the registers. The service equipment is usually represented by the central processing unit, and sometimes by the input equipment. The waiting discipline adopted is usually imposed by the supervisor program.

Clearly these interpretations are not the only possible ones, and in any case depend on the installation of which the queueing system is a model.

3. APPLICATION OF THE SIMPLEST MODELS TO DESCRIPTION OF THE ELEMENTS OF DATA PROCESSING SYSTEMS

The model of an isolated queueing system presented above is often well suited to describing the operation of the individual elements of an information system. Adaptation of some results of queueing theory, for instance, makes it possible to determine times of waiting for the beginning of transmission in a computer, minimum buffer capacities, the throughput of information channels, etc.

What follows is the most important results obtained from investigation of the elements of information systems. We shall limit ourselves in this discussion to presenting the effects of waiting time for service and of the distribution of queue lengths, since these are the parameters most often used in applications.

We introduce the following designations:

Let $G(x) = P\{\tau_i \leqslant x\}$ be the common distributor of values of τ_i independent in the set. Let us use the symbol η_i for the service time for each problem, and the symbols $H(x) = P\{\eta_i \leqslant x\}$ for the distributor corresponding to this time.

In the simplest queueing system (a so-called M/M/m system, in which the first symbol designates the type of input stream, the second symbol the type of service, and the third the number of channels) it is assumed that the input stream is a Poisson stream, i.e., that

$$G(x) = 1-\exp(-\lambda x),$$

and the service time η has an exponential distribution

$$H(x) = 1-\exp(-\mu x).$$

In these formulas λ and μ represent the arrival rate of problems and the service rate, respectively.

Let p_k represent the probability that there are k problems in a steady state in a queueing system, and let γ be the waiting time for service.

For a M/M/m system with FIFO service Erlang's classic formulas apply:

$$p_k = \frac{\varrho^k}{k!} p_0 \quad \text{at} \quad 1 \leqslant k < m,$$

$$p_k = \frac{\varrho^k}{m! m^{k-m}} p_0 \quad \text{at} \quad k \geqslant m, \tag{1}$$

where

$$p_0 = \left(\sum_{k=0}^{m} \frac{\varrho^k}{k!} + \frac{\varrho^{m+1}}{m!(m-\varrho)} \right)^{-1}, \quad \varrho = \frac{\lambda}{\mu}.$$

Waiting time for the beginning of service is covered by the formula

$$P\{\gamma \leqslant x\} = 1 - \left(\sum_{k=m}^{\infty} p_k \right) \exp[-(m\mu - \lambda)x]. \tag{2}$$

For a queueing system with a Poisson input stream and a general distribution of service time (a so-called M/G/1 system) Chinczyn's formula applies to the Laplace-Stieltjes transform of the distributor of waiting time for the beginning of service (cf., for instance, [1]):

$$\Phi(s) = \frac{1 - \lambda\tau}{1 - \lambda \dfrac{1 - h(s)}{s}}, \tag{3}$$

where

$$\tau = \int_0^{\infty} x \, dH(x),$$

$$h(s) = \int_0^{\infty} \exp(-sx) \, dH(x).$$

A steady solution exists for $\lambda\tau < 1$.

If as before p_k denotes the probability that there are k problems waiting in a queueing system in a steady state, then the Pollaczek-Chinczyn formula holds for the function of the growing number of problems waiting in the system:

$$\sum_{k=0}^{\infty} p_n z^n = \frac{(1-\varrho)(1-z)h[\lambda(1-z)]}{h[\lambda(1-z)] - z}. \tag{4}$$

In the work now being done on application of queueing theory to the investigation of information systems, results concerning more general queueing systems (such as those with general input streams) are rarely used and we shall not consider them.

As noted previously, the above results can be applied to the investigation of individual elements in data processing systems. An example of such an application is the very interesting analysis of disk operation (the IBM 2314) described in [2] (see also [16]). The model considered there

reduced to type M/G/1 systems operating in tandem. The analytic results were obtained on the basis of Chinczyn's formula given above.

An isolated queueing system with FIFO discipline is too simple a model to be useful in describing real information systems in general. Consideration of features such as multi-access, real-time operation, time sharing, and the like requires the construction of a queueing model. This effort has taken two basically different directions. The first direction deals with individual queueing systems with complicated priority systems. The other is connected with the investigation of certain special queueing systems, so-called feedback systems. We shall now consider the current state of work in both directions.

4. PRIORITY SYSTEMS

A feature of the majority of systems operating in real time is a highly developed system of priorities. Models of such systems are queueing systems with usually isolated service equipment, many input streams, and appropriate service discipline distinguished by a priority system. Several classes of priorities are recognized with their corresponding service disciplines. The appearance in the system of a problem with a priority higher than that being serviced may cause immediate suspension of service. This is called preemptive service discipline.

The problem whose servicing was interrupted may then be variously treated:

(1) Service is taken up from the point of interruption (preemptive-resume discipline);

(2) Service is repeated from the beginning at the original rate (preemptive--repeat-identical discipline);

(3) Service is repeated from the beginning at a different rate (preemptive--repeat-different discipline).

If the arrival of a problem with a higher priority does not cause suspension of service on a lower-priority problem, one speaks of non-preemptive discipline.

A finite number of priority classes is usually assumed. And the problems belonging to a given priority class are usually serviced according to FIFO discipline.

In priority systems a basic parameter which is investigated is the time necessary for complete servicing of a problem belonging to class k. This time is the sum of the waiting time for service and the so-called completion time, calculated from the moment of the beginning of service to the moment it has been completed.

A detailed analysis of type M/G/1 systems with priorities and preemptive discipline is given in the article by Wei Chang [3].

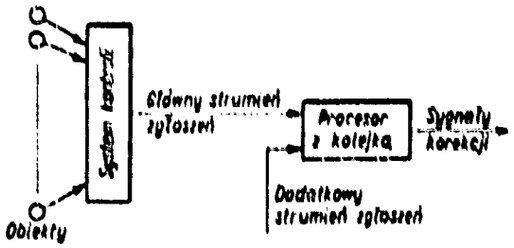

Figure 2. Fragment of a control system for aircraft flight (see example).

Obiekty = objects, or industrial processes
System kontroli = control system
Glowny strumien zgloszen = main problem stream
Procesor z kolejka = processor with queue
Dodatkowy strumien zgloszen = secondary problem input stream
Sytnaly korekcji = correction signals

By way of illustration, let us consider the application of the results given there to analysis of a simple system (Figure 2), which controls several objects when they fail to operate properly. The model described here is a modification of a fragment of a system for aircraft flight control, the conception of which is given in [6]. The arriving problems are signals obtained from the control system, indicating excessive deviation of the parameters of the objects from the planned parameters. Let us assume that the input stream of problems can be described with sufficient accuracy as a Poisson distribution at a rate λ_1. All these problems belong to a single priority class ($k = 1$). Service by the processor consists in computing correction signals for the incorrect parameters. Let $H_1(x)$ designate a distributor of the time necessary to service each problem.

In addition to executing correction programs the processor performs a number of other tasks. Requests for these services make up another Poisson stream of problems arriving at a rate λ_2. The problems arriving in the second stream have a lower priority ($k = 2$) and their service time is characterized by the distributor $H_2(x)$. For purposes of simplification we also assume that the source of problems in both streams is infinite, that the service time is independent of the number of problems in the system, and that there is no limitation on queue length.

Problems arriving in the main and the secondary streams form a queue with preemptive-repeat service discipline (i.e., type 1).

The basic parameters which we want to find are the time passing from the moment of arrival of a signal indicating improper functioning of the

object to the moment of computation of the correction signals and the time for executing secondary programs. We shall designate the distributors of these times as w(r) and v(t) respectively (we are considering only steady operation of the system). Computation of these distributors is easy if we know the distributions of times for waiting for beginning of service and for completion of the solution.

Let us use the symbols $W_1(s)$ and $W_2(s)$, to denote the Laplace-Stieltjes transforms of these waiting times, and $C_1(s)$ and $C_2(s)$ for the transforms of times for completion of solutions.

In the case of the main input stream an expression for $W_1(s)$ is obtained directly from the above-mentioned formula of Chinchyn (3):

$$W_1(s) = \frac{1-\lambda_1 \tau_1}{1-\lambda_1 \dfrac{1-h_1(s)}{s}},$$

where

$$\tau_1 = \int_0^\infty x\,dH_1(x),$$

$$h_1(s) = \int_0^\infty \exp(-sx)\,dH_1(x).$$

The time required for completion of service in the case of the main input stream equals the required service time, i.e.

$$C_1(s) = H_1(s).$$

The total time necessary for computing correction signals (from the moment the signal arrives indicating improper functioning to the moment when the computation is completed) thus has a distribution defined as follows:

$$\begin{aligned} w(t) &= w_1(t) * c_1(t), \\ w_1(t) &= \mathcal{L}^{-1} W_1(s), \\ c_1(t) &= \mathcal{L}^{-1} C_1(s). \end{aligned} \tag{5}$$

For the secondary stream the relationships are much more complicated:

$$C_2(s) = h_2\{s+\lambda_1[1-\Gamma_1(s)]\}, \tag{6}$$

where $\Gamma_1(s)$ is a root of the equation

$$z = h_1\{s+\lambda_1(1-z)\},$$

and $h_2(u)$, as before, is given by the formula for the Laplace-Stieltjes transform:

$$h_2(u) = \int_0^\infty \exp(-ux)\,dH_2(x).$$

The following equation is found for the waiting time for the beginning of service to problems arriving in the secondary stream

$$W_2(s) = \frac{1-\lambda_1\tau_1}{1-\lambda_1\dfrac{1-C_2(s)}{s}} \cdot \frac{s+\lambda_1[1-H_1^*(s)]}{(1-\lambda_1\delta_1)s}, \tag{7}$$

in which

$$\delta_1 = \frac{\tau_1}{1-\lambda_1\tau_1}, \quad \tau_2 = \int_0^\infty x\,dH_2(x).$$

The time for completing a program arriving in the secondary stream [is found] in terms of a distributor expressed by $w_2(t) * c_2(t)$.

If the limitations on service time for problems arriving in the main and the secondary streams are known, the above expressions make it possible to formulate the required processor parameters.

For a system such as the ones described the results of queueing theory permit computing the length of both types of queues. This makes it possible to evaluate the parameters of the system being planned, such as the size of memory associated with the processor, the way in which it is divided up, and the like.

In a number of cases the requirements placed on the system are confined to placing limitations on the moments of certain distributions (waiting time, time for completing programs, queue length). Thus there is no need to compute the reverse Laplace transforms, and the moments themselves are found by differentiating the expressions for the transforms.

Let

$$w_1^{(r)} = \int_0^\infty x^r\,dw_1(x), \quad r = 1, 2, \ldots,$$

$$\tau_1^{(r)} = \int_0^\infty x^r\,dH_1(x), \quad r = 2, 3, \ldots.$$

As we know

$$w_1^{(r)} = (-1)^r \frac{d^r W_1(s)}{(ds)^r}\bigg|_{s=0}. \tag{8}$$

For the first two moments we find

$$w_1^{(1)} = \frac{\lambda_1\tau_1^{(2)}}{2(1-\lambda_1\tau_1)}, \tag{9}$$

$$w_1^{(2)} = \frac{\lambda_1\tau_1^{(3)}}{3(1-\lambda_1\tau_1)} + \frac{\lambda_1^2(\tau_1^{(2)})^2}{2(1-\lambda_1\tau_1)^2}.$$

If, for example, we know the limitations on the value of the mean time required to compute an aircraft flight correction

$$w_1^{(1)} + \tau_1 \leqslant a$$

and if we use the symbol v to denote the speed of the processor which we are attempting to determine (i.e., in operations per unit time), and the length of the program for computing the correction is constant for all flights and equals l (operations); in other words,

$$H(x) = \begin{cases} 0, \text{ where } x < \frac{l}{v}, \\ 1, \text{ where } x \geq \frac{l}{v}, \end{cases}$$

the following inequality is obtained

$$\frac{\lambda_1 (l/v)^2}{2[1 - \lambda_1 (l/v)]} + \frac{l}{v} \leq a.$$

When, for example, $\lambda_1 = 0{,}1$ and $a = 5$ we obtain $(l/v) \leq 3{,}8$; if the length of the correction program is known we can use this value to determine the required speed of the computer.

In real systems it happens that the input stream cannot be satisfactorily described by a Poisson distribution. This causes considerable analytical difficulties, and is responsible for the fact that the literature lacks satisfactory results for the corresponding queueing models with priorities and preemptive service discipline.

Developments in so-called mixed-priority queueing systems are interesting for applications to real data-processing systems [4, 5]. In this type of system the currently serviced problem (e.g., the program being executed) can create a signal which disables suspension of service until another signal enables suspension. As a result the total service time may be made up of arbitrarily located time segments of three types: a non-preemptive segment, a preemptive segment with resume discipline, and a preemptive segment with repeat discipline. Figure 3 shows this type of service.

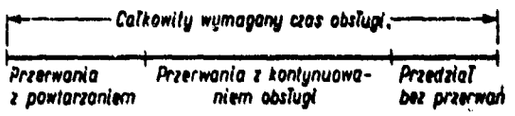

Figure 3. Example of breakdown of service time segments

Calkowity wymagany czas obslugi = total service time required
Przerwania z powtarzaniem = preemptive-repeat segment
Przerwania z kontynouwaniem obslugi = preemptive-resume segment
Przdzial bez przerwan = time segment without suspension

Further development and implementation of this type of model assumes that a problem whose service was preempted requires before each resumption of service a special set-up time period to recreate the conditions prevailing before preemption.

A detailed analysis of M/G/1 systems with mixed priorities is contained in L. Schrage's article [5]; the conclusions are quite complicated and we shall not cite them here.

5. QUEUEING SYSTEMS WITH FEEDBACK AS MODELS FOR MULTI-ACCESS SYSTEMS

The application of queueing theory to the investigation of multi-access systems with feedback makes possible the development of interesting models of such systems.

Typical multi-access systems operate on the so-called time quantum principle. A model of this type of system is a model of the distribution among users of a resource, which is the working time of the central processing unit.

Let us consider the simplest of the models of this type, the so-called RR (round-robin) model shown in Figure 4.

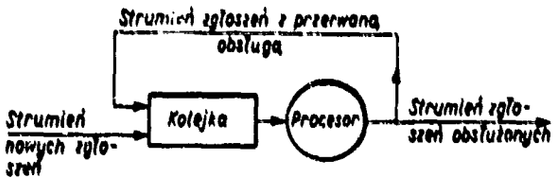

Figure 4. Model of an RR system

Strumien nowych zgloszen = input stream with new problems
Kolejka = queue
Procesor = central processing unit
Strumien zgloszen z przerwana obsluga = stream of suspended problems
Strumien zgloszen obsluzonych = output stream of completed problems

New problems (programs) entering the RR system go to the end of the processor queue. Problems are serviced in the order in which they enter the queue. Each problem obtains a quantum q of time from the processor. If the problem is completed before the end of the service time, the problem leaves the system. Otherwise service is suspended and the problem goes back to the end of the queue.

In the simplest RR system it is assumed that the time necessary to send a problem to the end of the queue equals zero. In addition it is assumed that the input stream is a Poisson stream at a rate λ and that it comes from an infinite source, and that the time necessary to service a problem is a random variable with an exponential distribution with a parameter μ.

Kleinrock [7] has investigated a number of interesting parameters of RR systems.

A particularly interesting model can be obtained by considering the limiting case, in which the quantum of service time allotted approaches zero ($q \to 0$). The service system which this produces is called a processor-sharing (PS) system. In a PS without priorities all problems are serviced simultaneously at a rate equal to μ/n, where n is the number of problems in the system at any one time. The probability P_n of servicing simultaneously n problems in a PS system in a steady state is expressed by Erlang's formula, which was noted earlier:

$$p_n = \left(\frac{\lambda}{\mu}\right)^n \left(1 - \frac{\lambda}{\mu}\right) \quad \text{for} \quad n = 0, 1, 2, \ldots$$

The basic value characterizing the operation of an RR or PS system is the distributor $w(t, x)$ of the time spent in the system by a problem requiring t time of the entire processor for its completion. Articles [7] and [8] contain formulas for computing the mean value $W(t)$ of the time spent by a problem in RR and PS systems. For a PS system

$$W(t) = \frac{t}{1 - (\lambda/\mu)}.$$

Comparison of this result with the corresponding result for a queueing system with FCFS service shows that a PS system prefers short programs: problems for which the required processor time is less than $1/\mu$, remain in a PS system for less time than they do in a queueing system with FCFS discipline.

The most recent working on PS systems [9] has found a formula for the Laplace-Stieltjes transform $V(t, s)$ of the distributor $v(t, x)$ of the difference between the time spent in the system and time t.

There are also more general conclusions (M. Sakata, cf. [9]) for a PS system assuming that the distribution of service time is general, with the reservation that they concern the unconditional distribution of time spent in the system.

RR and PS models offer the possibility of introducing priorities [7, 8, 10]. In the case of an RR model a problem belonging to class p obtains g_p quanta of service time each time. In processor-sharing systems problems belonging to class p are services at a rate corresponding to the f_p part of the computing power of the processor:

$$f_p = \sum_{i=1}^{p} \frac{g_p}{g_i m_i},$$

where P is the number of classes, and m_i is the number of problems in class i currently in the system.

The necessary formulas for mean time in the system are given in [7].

In multi-access systems with time sharing the privileges accorded to short programs by the service discipline corresponding to the RR model are

often insufficient. In the operation of such a system there are usually periods of increased rate of arrival of problems, which causes overloading. In that event long programs are particularly inefficient. Repeated return of such programs to the processor blocks up the system. The solution employed is to use a service discipline corresponding to the queueing model with several queues and feedback. The literature considers several such models.

FB_N models (systems without priorities). The model has N queues. Problems enter the system and are placed at the end of queue number 1. Servicing of problems in this queue is subject to FCFS discipline and the time quantum principle. The processor (at least in an RR system) allots each problem a quantum q of service time. Problems requiring no more than q time to reach completion leave the system finished. Longer problems go to queue number 2. In queue number i $(1 < i < N)$ problems are services according to the FCFS principle. A problem obtains its quantum q of service time and if this time was sufficient to complete the computation, it leaves the system; otherwise it proceeds to the end of queue number (i + 1). The processor begins to service problems in queue i only when all the queues with numbers less than i are empty. Problems in queue number N are serviced according to the FCFS principle to the very end, except that after each time quantum service may be suspended if a non-empty queue is found with a number less than N.

The case of the limiting FB_N system as $q \to 0$ is not interesting, since the operation of such a system when $N < \infty$ reduces to the operation of an individual queueing system with FIFO discipline. More interesting is an FB_N system when $N = \infty$ and as $q \to 0$.

FB_N systems without priorities and with Poisson input streams and exponential distribution of service time have been thoroughly examined by a number of authors [7, 8, 11]. Schrage [13] has made more general consideration of FB_N systems when $N = \infty$ with general distribution of service time. These articles also consider systems in which the source of problems has a finite capacity, as well as of systems with an infinite source.

These generalizations of FB_N systems consist in considering systems with priorities and with various values for the quantum of time allotted problems waiting in the various queues.

In priority systems the input stream is divided and problems are directed to different queues, depending on the class to which they belong. Service in individual queues is also subject to priority discipline, with the highest-priority queue being the one soonest serviced.

FB_N models with priorities correspond to systems in which program privileging is gradually introduced, depending on parameters such as length of program and amount of memory used. The most important results for FB_N systems with priorities are given in [7] and [14].

The RR_r systems examined in [15] are a variant of FB_N systems. The difference between RR_r systems and FB_N systems lies in the different service discipline applied to problems in the highest-numbered queue r. The last queue in an RR_r system is an RR system. As a result the extension of execution times for programs requiring the longest processor time is greater than in an FB_N system. The proposed service discipline for problems corresponding to an RR_r system (as in the case of FB_N systems) is intended to provide for automatically shifting the execution of the longest programs to a period of low problem input.

The multi-access systems actually implemented correspond both to the simplest models described above and to more complex ones [11]. For example, the GE-Darmouth Time-Shared System, which serves 200 users, employs the simplest RR algorithm with q = 200 [ms?]. The MAC system developed at MIT is a modification of an FB_9 system with different time quanta in different queues ($q_n = 2^{n-1} \cdot 0{,}5$ s, where n is the queue number).

Some systems are attempting to employ several different service algorithms; the SDC Time-Sharing System, for instance, uses an RR algorithm with q = 400 ms and a modification of an FB_2 system.

Investigation of models of multi-access systems with time sharing are at present directed more toward analysis than toward synthesis of systems. The assumptions under which analytic results are obtained (special stream distributions, stochastic independence of streams, particular forms of distribution of service time) can under real conditions become too severe limitations. There is a lack of results at present on the operation of systems in steady states. Not very much work has been done in analyzing the losses associated with suspending the execution of programs. It should be stressed, however, that analytic methods have certain advantages over modeling methods: it is easy to analyze the influence of parameters, and there is no need to introduce simulator programs, which can be disrupting.

6. CONCLUSION

The problems discussed concern only a few applications of the apparatus of queueing theory in information systems. A great deal of research concerns information transmission systems, for example. In this area application of the methods of queueing theory permits computation of the throughput of complicated communications networks (cf. [16] and [17]). Some results concerning queueing networks [18] can be applied to research on complex information systems.

BIBLIOGRAPHY

1. Gnedenko, B. V. and Kovalenko, I. N.: Vvedeniye v teoriyu massovogo obsluzhivaniya [Introduction to Queueing Theory], Moscow, 1966, published by Nauka.

2. Abate, J., Dubner, H., and Weingerg [sic], S.B.: Queueing analysis of the IBM 2314 disk storage facility, J. ACM, 1968, Vol 15, No 4.

3. Chang, W.: Preemptive priority queues, Operations Res., 1965, No 5.

4. Chang, W.: Queueing with nonpreemptive and preemptive resume priorities, Operations Res., 1968, No 6.

5. Schrage, L.E.: Mixed priority queues with applications to the analysis of real-time systems, Operations Res., 1969, No 4.

6. Libura, M. and Walukiewicz, S.: General principles and data-processing algorithm for a system for civil aircraft flight control, Warsaw, 1969, Publications of the Institute of Automation of the Polish Academy of Sciences No 83.

7. Kleinrock, L.: Time-shared systems, A theoretical treatment, J. ACM, 1966, No 3.

8. Cofman, E.G., and Kleinrock, L.: Feedback queueing models for time-shared systems, J. ACM, 1968, No 4.

9. Cofman, E.G., Muntz, R.R., and Trotter, H.: Waiting time distributions for processor-sharing systems. J. ACM, 1970, No 1.

10. Rasch, P. J.: A queueing theory study of round-robin scheduling of time-shared computer systems, J. ACM, 1970, No 1.

11. Estrin, G.: Measures, models, and measurements for time-shared computer utilities, Proc. 22nd Nat. Conf. ACM, Thompson Book Co., Washington, D.C., Academic Press, London, 1967.

12. On-line Computing, New York, 1967, McGraw-Hill.

13. Schrage, L.E.: The queue M/G/1 with feedback to lower-priority queues, Management Sci., 1967.

14. Cofman, E.G.: Stochastic models of multiple and time-shared computer operation, Rep. No. 66-38, Dept. of Engineering, University of California at Los Angeles, 1966.

15. Adri, J., Avi-Itzhak, B.: A time-sharing model with many queues, Operations Res., 1969, No 6.

16. Massovoye obsluzhivaniye v sistemakh peredachi informatsii [Queueing in information-transmission systems], Moscow, 1969, published by Nauka.

17. Vasharin, T. P., Kharkevich, A.D., and Shneps, M.A.: Massovoye obsluzhivaniye v telefonii [Queueing in Telephony], Moscow, 1968, published by Nauka.

18. Posner, M., and Bernoltz, B.: Closed fine queueing networks with time lags and with several classes of units, Operations Res., 1968, No 5.

Received for publication 1 June 1970.