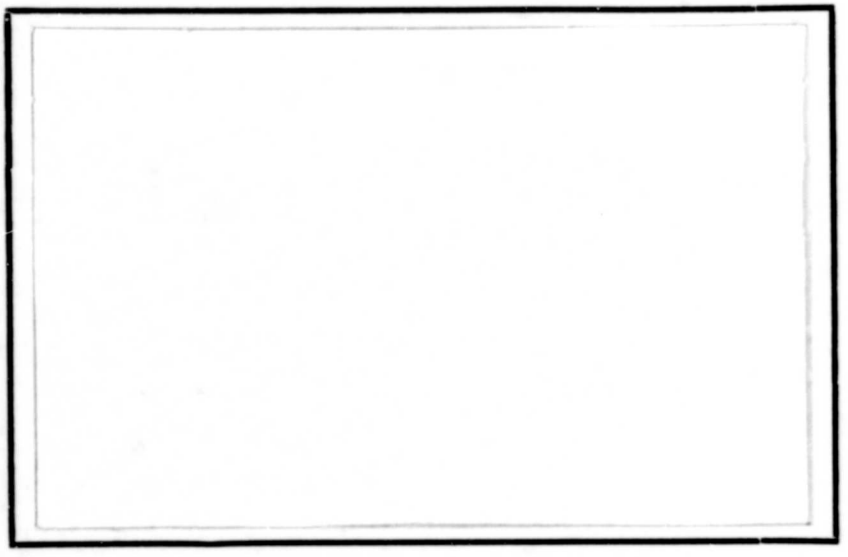

הטכניון — מכון טכנולוגי לישראל

הפקולטה להנדסת תעשיה וניהול

TECHNION — ISRAEL INSTITUTE OF TECHNOLOGY

FACULTY OF INDUSTRIAL AND MANAGEMENT ENGINEERING

HAIFA, ISRAEL

QUEUEING MODELS FOR TIME-SHARING SERVICE SYSTEMS

by

I. ADIRI and B. AVI-ITZHAK

Operations Research, Statistics and Economics
Mimeograph Series No. 27

This research has been sponsored in part by the Logistics and Mathematical Statistics Branch, Office of Naval Research, Washington, D.C., under Contract N61052-68-C-0014.

October 1968

ABSTRACT

In most queueing situations, it is desirable that service to customers be free from interruptions causing time and service losses. It is recognized, however, that in certain circumstances controlled interruptions may improve overall system performance, and this idea is the basis of all time-sharing systems. In the latter, service is given in segments; the customer may be served, uninterrupted, for no longer than some predetermined time interval called a quantum. If the service is not completed within a given quantum, the customer is dismissed and placed in a queue, and the next customer is admitted.

There are two main advantages to such queueing disciplines: (a) the expected waiting time of a customer is an increasing function of the amount of service demanded by him, and (b) the total value of the service is sometimes increased by breaking it down into segments. This is especially true when running certain types of computer programs. The obvious disadvantages of the time-sharing models are set-up time losses and sophisticated supervision and coordination requirements.

This paper is a survey of time-sharing models studied recently by the authors. It covers the following single-server models: (i) single queue with infinite number of potential customers - R.R.1; (ii) single queue with finite number of potential customers; (iii) r queues with infinite number of potential customers - R.R.r, in which a customer who completes his i-th service segment joins the end of the (i+1)-th queue, the r-th queue is organized on a "round-robin" basis, and the server, when admitting a customer to service, selects the first in the lowest-index's non-empty queue; (iv) R.R.1. with various types of priority regimes.

The main quantities obtained in explicit form are queue sizes, overall unconditional waiting times, and waiting times conditional on the length of service demanded by a given customer. Performance optimization with respect to quantum size is also considered.

Each model provides a means for achieving desired given properties. The performance parameters of the models are compared numerically, and the advantages and weaknesses of each model are discussed.

QUEUEING MODELS FOR TIME-SHARING SERVICE SYSTEMS

I. Adiri and B. Avi-Itzhak

INTRODUCTION

Time-sharing is a means for providing service to a large and diversified population of customers, using remote-access devices, with a view to ensuring shorter response times to short service demands at the expense of the long ones and improving customer-server interaction. The queue discipline commonly used in single server time-sharing systems is known as the round-robin schedule (abbreviated R.R.). Under it, service is given in segments called quanta, and customers go in and out of service in a cycle, receiving one quantum in each cycle until service is completed. In the simplest form of this type of schedule, called R.R.1, there is only one queue. Having completed a quantum of service, the server admits the first customer in the queue, who is in turn entitled to one quantum at the end of which he either has completed his service and departs, or else joins the end of the queue. Newly arrived customers also join the end of the queue.

There are two main advantages to such a queue discipline: (a) The expected waiting time of a customer is an increasing function of the amount of service demanded by him; this effect is achieved without prior knowledge of the specific service length demanded by an arriving customer, and where these service times are known, the same kind of effect is obtainable by using "shorter-service-first" type disciplines. (b) The service is given in small segments. with the customer waiting between two consecutive segments. Suppose the total value derived by a customer is the same whether he receives his service continuously or in segments. In these circumstances, in the case of a time-sharing queue, the customer receives some of his value earlier, assuming that each segment has some value. Sometimes the total value of the service is increased by breaking it down into segments; this is especially true when running certain types of computer programs, where knowledge of partial results may help determine how to continue the run, if at all. This property of time-sharing computer systems is called the "conversational mode". If the quantum is short, the customer

may gain the impression that he is the only user of the computer, since he has an individual remote console connected to it and is unaware of the other users.

In practice, there are obvious disadvantages to the time-sharing queueing discipline. The server is compelled to spend time supervising the traffic and preventing possible service losses due to interruptions. For example, in a computer system there are "housekeeping" and "overhead" time losses, and in addition the supervising program is complicated and occupies a larger portion of the available magnetic core memory. Thus, only a certain fraction of each quantum is utilized for service and the remainder may be regarded as losses.

This paper reviews a number of time-sharing queueing models recently studied by the authors, with a view to possible applications to computer systems with remote access. The models considered here have a number of features in common. In particular, it is throughout that customers arrive at a single server system according to a Poisson process and that service requirements are mutually independent and exponentially distributed with mean $1/\mu$. The quantum comprises two parts, the first being assumed to be of constant length τ (the set-up time of the server representing losses due to the use of the R.R. schedule) and the second, of maximum length θ, known as the quantum processing time, devoted to service. The actual length of a quantum is, therefore, a random variable taking values in the interval $[\tau, \tau + \theta]$.

Model I - R.R.1

The simplest model of the type under study is the R.R.1 model where there is but one queue and the population of potential customers is assumed to be infintely large, this being mathematically reflected in the assumption that customers arrive according to a homogeneous Poisson process with intensity λ. The exact queue discipline of this model has been described in the introduction. Denoting the length of the quantum by Q, we have:

$$Q \overset{d}{=} \begin{cases} \tau + \theta & \text{with probability } \alpha = e^{-\mu\theta} \\ D & \text{" } \quad \text{" } \quad 1 - \alpha \end{cases}, \tag{1}$$

$\overset{d}{=}$ denoting equal in distribution, and D having a density $f_D(.)$, and

$$f_D(x) = \begin{cases} \mu e^{-\mu(x-\tau)}/(1-\alpha) , & \tau \leq x \leq \theta+\tau , \\ 0 , & \text{otherwise} . \end{cases} \tag{2}$$

The total time devoted by the server to a single customer, including set-up times, is denoted by S

$$S = (R-1)(r+\theta) + D , \tag{3}$$

where R is the number of service quanta given to the customer and is geometrically distributed. The Laplace-Stieltjes (L.S.) transform of S is readily obtained as:

$$L_S(z) = \frac{\mu e^{-z\tau}(1-\alpha e^{-\theta z})}{(\mu+z)(1-\alpha e^{-z(\tau+\theta)})} . \tag{4}$$

Let X_n be the number of customers in the system immediately after the n-th departure. Then

$$X_{n+1} = X_n + Y - \delta$$

where Y is distributed as the number of arrivals falling in a time interval of length S, and

$$\delta = \begin{cases} 0 & \text{if } X_n = 0 \\ 1 & \text{if } X_n > 0 \end{cases} \tag{5}$$

From (5) it is obvious that the Markov chain $\{X_n\}$ is statically the same as in a M/G/1 model where service times are distributed as S[*]. The condition for $\{X_n\}$ to be positive recurrent is

$$\rho = \lambda E(S) < 1 \tag{6}$$

Suppose X_n is stationary and let X be distributed as X_n, n=1,2,...,. The r.v. X represents the steady state number of customers present in the system at points of departure. From the M/G/1 model we have the generating function of X as

$$G_X(z) = \frac{(1-\rho)(z-1)L_S(\lambda(1-z))}{z - L_S(\lambda(1-z))} , \tag{7}$$

[*]This property has been already recognized by Takács[4].

and

$$E(X) = \rho + \frac{\lambda^2 E(S^2)}{2(1-\rho)} \quad . \tag{8}$$

It should be pointed out that the distribution of X is the same as the stationary distribution of the number of customers present in the system at a point in time chosen at random. This property is a result of the arrival process being a homogeneous Poisson process.

We define the stationary first response time of a customer, denoted by T_1, as the time elapsing from his arrival until his first admittance to service. The stationary i-th ($i>1$) response time, T_i, is the time elapsing between the (i-1)-th and the i-th admittances to service of a customer demanding at least i service quanta. The total response time, T, which is the time a customer spends in the system, is made up of the sum of his partial response times. The expectation of T is obtainable from Eq. (8) using Little's[2] theorem.

A most meaningful measure in a time-sharing system is the total response time of a customer whose service demand, L, equals ℓ, the expectation of which is given as

$$E(T|L=\ell) = \sum_{i=1}^{\left[\frac{\ell}{Q}\right]} E(T_i) + \ell - Q\left(\left[\frac{\ell}{Q}\right]-1\right) + \tau \quad , \tag{9}$$

where $[x]$ is the smallest integer exceeding or equal to x. The expectations of T_i (as well as the L.S. transforms) were obtained in closed form after some rather tedious analysis, presented in detail in an earlier paper by the authors[3]. Substitution of $E(T_i)$ in Eq. (9) yields

$$\begin{aligned} E(T|L=\ell) = {} & E(Q)(E(X)-\rho) + \rho\frac{E(Q^2)}{2E(Q)} + \frac{1}{(1-\alpha)(1-\rho)}\Big\{\left(\left[\tfrac{\ell}{Q}\right]-1\right)(Q+\tau) \\ & ((1-\rho)(1-\alpha) + \lambda E(Q)) + E(Q)(1-(\rho+\alpha)(1-\rho))^{\left[\frac{\ell}{Q}\right]-1}) \\ & ((E(X)-\rho)(\alpha+\rho(1-\alpha)) + \frac{\rho}{2E(Q)}(\lambda E(Q^2) + 2\alpha(Q+\tau)) - \\ & - \frac{\lambda(Q+\tau)(\rho+\alpha(1-\rho))}{(1-\alpha)(1-\rho)}\Big)\Big\} + \ell - Q\left(\left[\tfrac{\ell}{Q}\right]-1\right) + \tau \quad . \end{aligned} \tag{10}$$

The influence of various system parameters on the performance of the system is illustrated graphically.

Figure 1 shows the relation between ρ and Θ for fixed values of μ and τ. As is expected, the time losses caused by the R.R.1 discipline are large when the value of Θ is small, and decrease with the Θ. It should be noted that $\rho > \lambda/\mu$ unless there are no set-up time losses, that is $\tau = 0$.

The expected waiting time of a customer whose service requirement equals ℓ, that is $E(T|L=\ell)$, is given in Figure 2 for various values of Θ. It is seen that $E(T|L=\ell)$ is practically a linear function of ℓ for almost all values of Θ. This implies that $E(T_1) \doteq E(T_2) \doteq \ldots E(T_j) \doteq \ldots$, hence Eq. (10), which is rather complex, can be replaced by an approximation based on the assumption that $E(T_i)$ are the same for all i.

$$E(T|L=\ell) \doteq \left([\tfrac{\ell}{\Theta}]-1\right) E(T_1) + \frac{\Theta}{2} + \tau \quad . \tag{11}$$

The explicit form of $E(T_i)$ is given in reference[3].

R.R.1 and FIFO are compared in Figure 3, which yields the service demand length for which the expected waiting times are equal in both disciplines. For example, if a customer requires 50 seconds of service and $\lambda/\mu = 0.8$, then for $\Theta \doteq 1.15$ his expected waiting time is the same in both disciplines. If his service demand is less than 50 seconds, he will be better off (on the average) in a R.R.1 discipline, and vice versa.

An important factor in designing a time-sharing system is the size of the quantum and its component parts, τ and Θ. In most cases τ is determined by the technological features of the system, and it is obviously desirable to keep it as small as possible. On the other hand, the magnitude of Θ is almost entirely up to the designer; hence the importance of defining our objective in terms of Θ, as the only controllable parameter of the R.R.1 system.

A cost function of the following form is used as the objective function to be minimized:

$$C(\Theta) = C_2 E(T_1) + \int_0^{\infty} C_1(\ell)(E(T|L=\ell)-\ell)\mu e^{-\mu\ell} d\ell \quad , \tag{12}$$

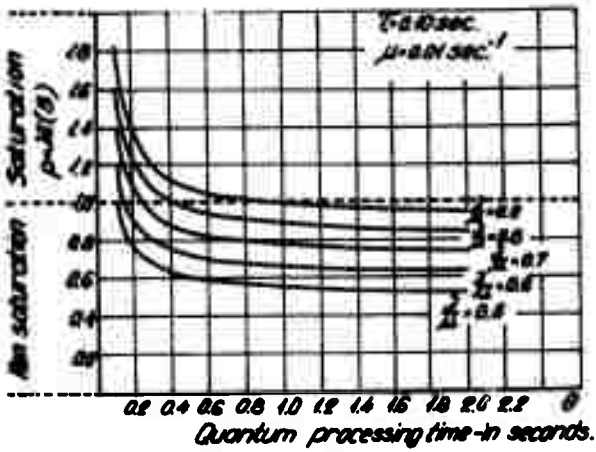

Figure 1: ρ as a function of θ for various values of λ.

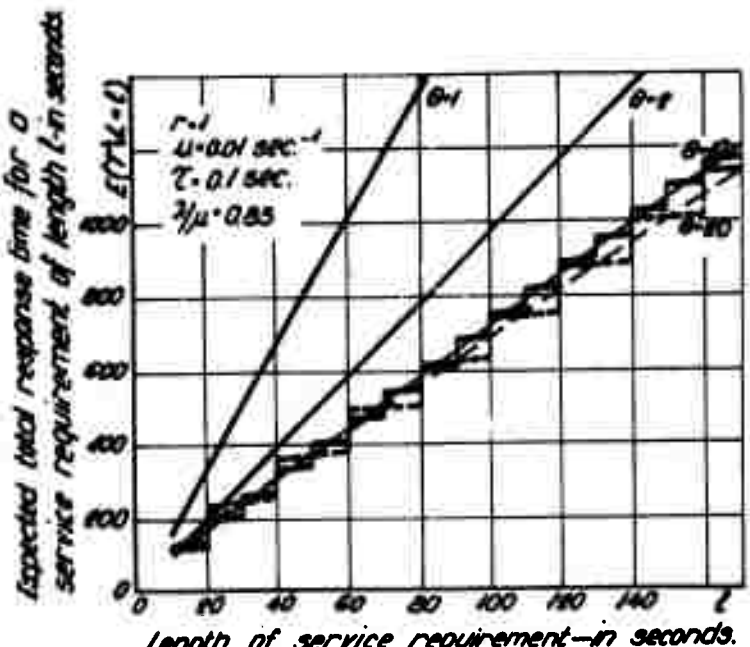

Figure 2: Expected Total Response Time for a Service Requirement of Length t as a function of t.
(For different values of θ, The R.R.1 Model).

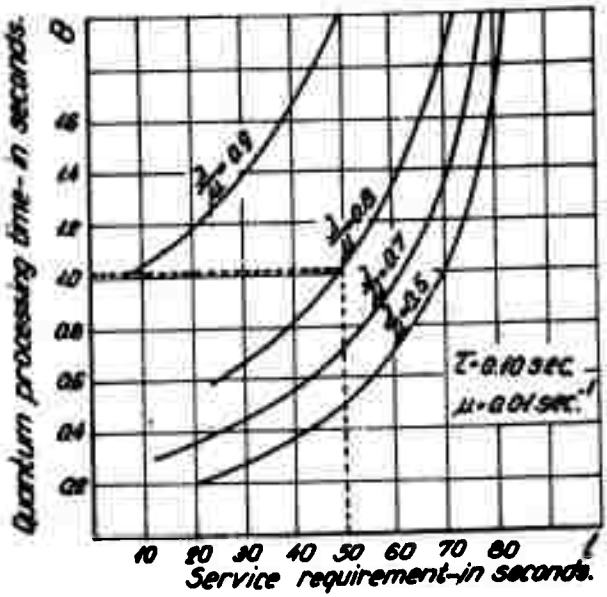

Figure 3: Comparison between R.R.1 and FIFO.

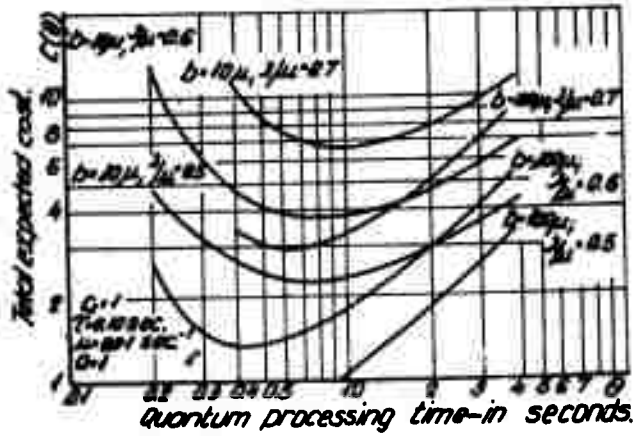

Figure 4: Total cost as a function of θ.

In this expression $C(\theta)$ is the average cost per customer. The first term on the right hand side of the expression is motivated by the desire to provide an arriving customer with ready access to the server; the second term represents the waiting cost. By choosing $C_1(\ell)$ as a decreasing function, we accentuate the priority given to short service demands. In Figure 4 the form of $C_1(\ell)$ is

$$C_1(\ell) = ae^{-b\ell} \quad . \tag{13}$$

Model II - R.R.1 Modified

In the preceding model it was assumed that even when there is only one customer in the system, he is served in quanta. This assumption may not be realistic, especially where λ/μ is small, since with the waiting line empty there is no reason for artificial interruption of service to a customer (first suggested by Schrage[4]), it is assumed accordingly that service may not be interrupted as long as the waiting line is empty. When there is a new arrival during service to such a customer, the latter may continue to be served for a period of length θ. If his service is completed within this additional period, he leaves the system, otherwise he is dismissed after θ time units and joins the end of the queue. In both cases, the server attends the next customer as soon as he becomes free. In all other respects the modified model is identical to the R.R.1. model.

The improvement in system performance due to the modification is obviously small when λ/μ is large. However, when λ/μ is small the improvement may be considerable. This is shown in Figure 5. The mathematical analysis of this model is given in reference[3].

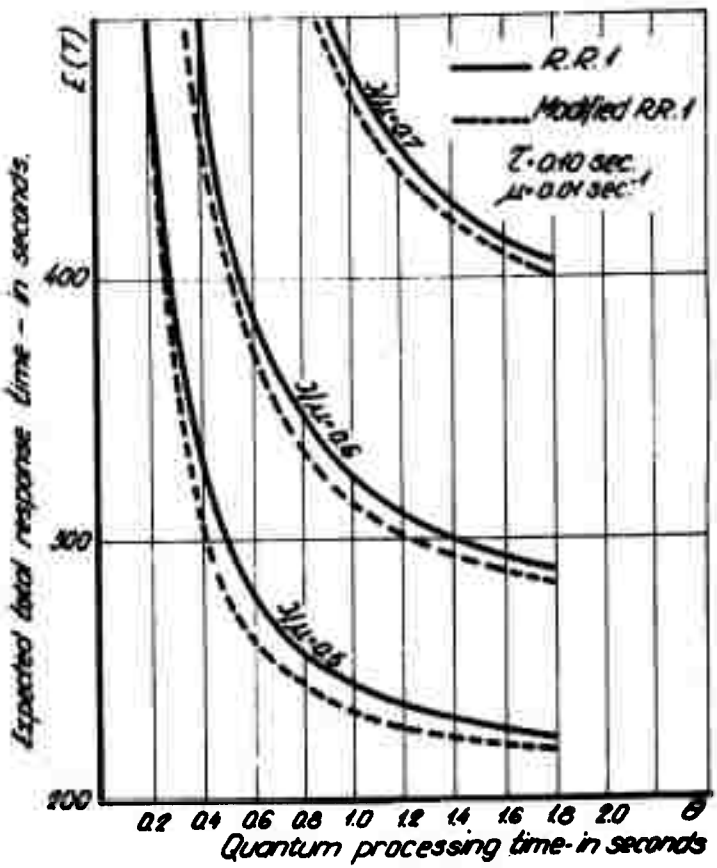

Figure 5: Expected total response times. A comparison between R.R.1. and modified R.R.1.

Model III - R.R.1 With Finite Number of Customers

In this model it is assumed that the number of potential customers is finite and equals N (corresponding to a computer system with a single control processor and N terminals), and that the time elapsing from the moment a customer completes his service until his next arrival at the queue is exponentially distributed with mean $1/\lambda$. In all other respects this model is identical to Model I (R.R.1). Detailed mathematical analysis is given in reference[5], based in part on the works of Takács[6] and Krishnamourthi and Wood[7].

Figures 6 and 7 show the expected total response time as a function of θ for different values of N. The case $N = \infty$, representing the R.R.1 model, is also plotted. The R.R.1 model serves as a good approximation when θ is large and the load, $N\lambda/u$, is low. Figures 8 and 9 show the relation between $E(T|L=\ell)$ and ℓ for different values of N. Here again it is seen that the approximation $N = \infty$ is improved as the load becomes lower. Note that the relations are practically linear for all illustrated cases.

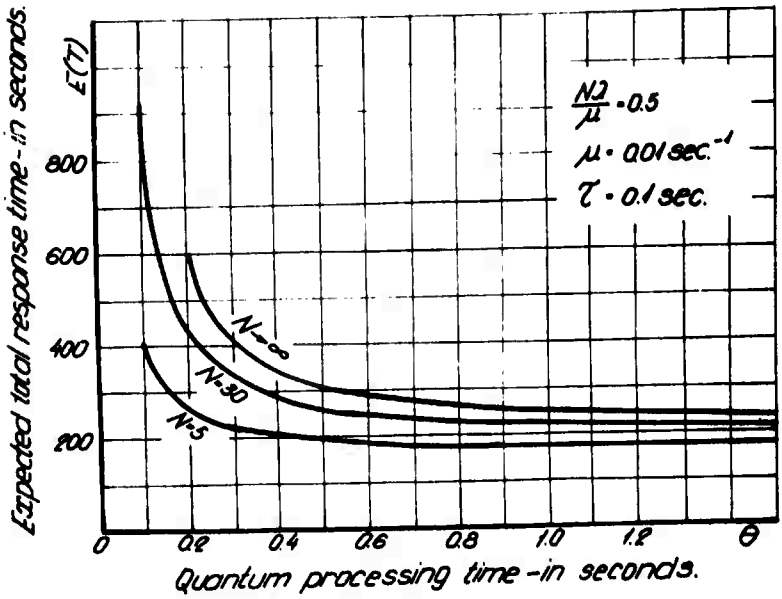

Figure 6: Expected Total Response Time as a Function of the Quantum Processing Time.

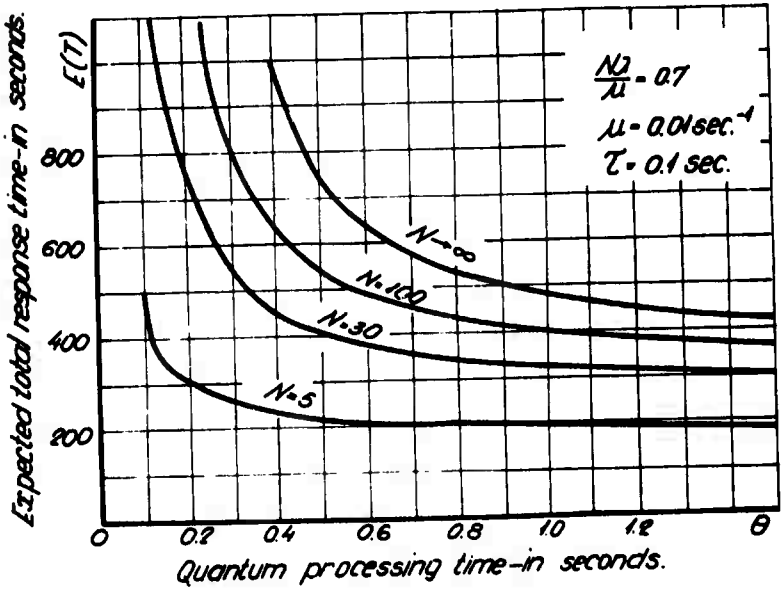

Figure 7: Expected Total Response Time as a Function of the Quantum Processing Time. .

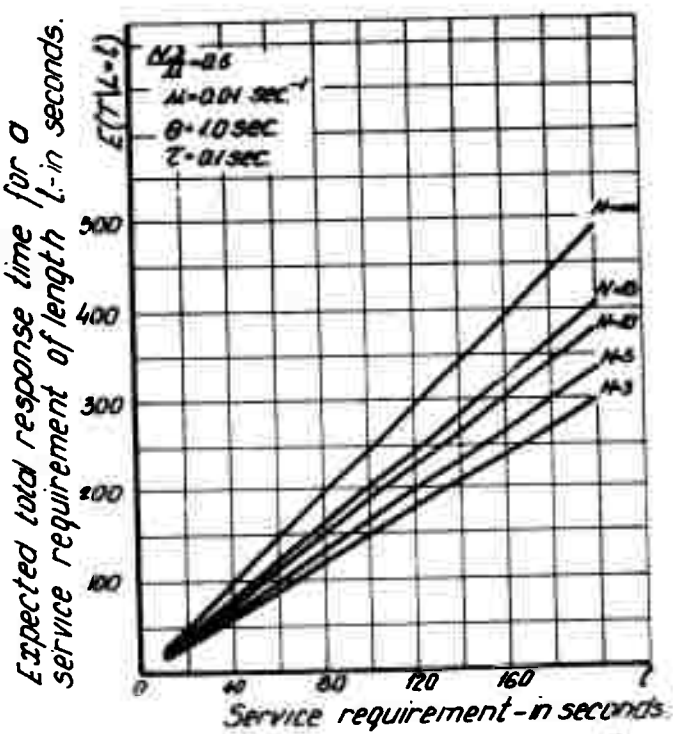

Figure 8: Expected Total Response Time for a Service Requirement of Length l as a Function of l.

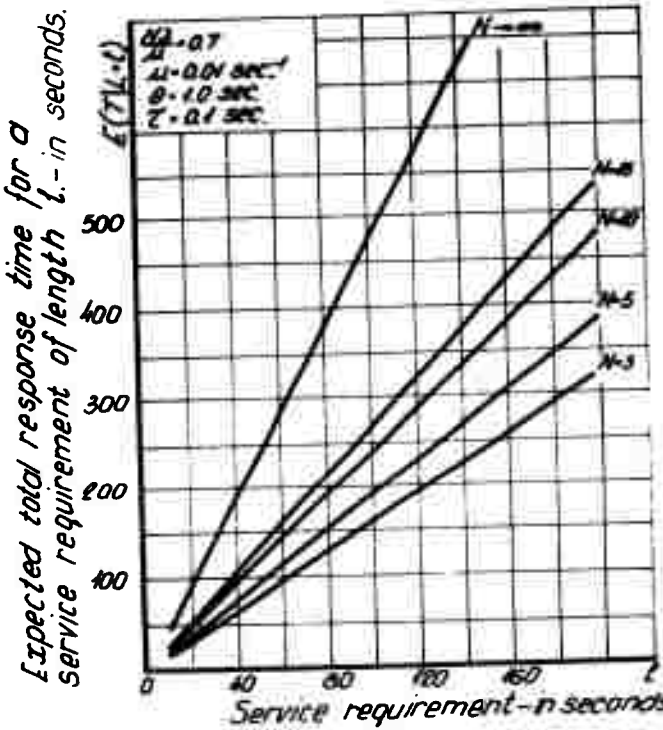

Figure 9: Expected Total Response Time for a Service Requirement of Length l as a Function of l.

Model IV - R.R.r

In all preceding models the server is ignorant of the amount of service (number of service segments or quanta) already given to the customers present in the system. Thus, it can be expected that the i-th response time of a customer is not strongly dependent on i, i.e., his expected total response time is approximately an increasing linear function of his service demand. This is demonstrated in Figures 2, 8 and 9.

One weakness of time-sharing systems, inherent in the R.R.1 model, is sensitivity to overloading by customers with long service demands (long processing times). The presence of customers with long service demands increases the waiting times of all customers in the system, and in cases of overloading (saturation) a demand cannot be satisfied in a finite time, however short it may be. On the other hand, there is no sufficient discouragement of customers with long service demands, since waiting time increases linearly with demand length. An actual time-sharing system is frequently overloaded during peak hours and underexploited during the rest of the day. It is desirable to transfer automatically long service demands, placed during peak hours to less loaded hours of the day, or even discourage them completely and assign them to batch processing during night hours. What is needed is, thus, a system with a following features: (i) waiting times of short demands less sensitive to the presence of long demands than in the simple round-robin discipline; (ii) short demands satisfied even under conditions of overloading (the higher the overloading, the shorter the demand must be in order to be satisfied); and (iii) effective discouragement of long demands during peak hours, by postponing their service to non-peak hours. Needless to say, the above must be realized without prior knowledge of demand lengths.

The R.R.r. model is proposed as a solution along the above lines. It comprises a single server and r queues. Customers arrive according to a homogeneous Poisson process with intensity λ. The waiting line consists of r separate queues. A newly arrived customer joins the end of the first queue; upon admittance to service he is eligible for θ_1 units of service time. If his service demand is shorter than θ_1, he completes his service and leaves the system, otherwise he receives θ_1 units of service time and joins the end of the second queue. In general, at the i-th queue

$(i=1,2,\ldots,r-1)$ the customer is eligible for θ_i units of service time during which he may complete his service and leave the system, or join the end of the $(i+1)$-th queue after receiving θ_i units of service time. A customer in the r-th queue who does not complete his service within θ_r service times units re-joins the end of the same queue. Having completed a service quantum at any one of the queues, the server admits to service the customer who is first in the lowest index non-empty queue. We assume that each quantum starts with a set-up time due to swapping, housekeeping, etc. The i-th quantum thus comprises two elements: a set-up time of length τ_i and a processing time not exceeding θ_i units.

The mathematical analysis is omitted here, and the interested reader is referred to an earlier paper by the same authors[8]. Here the main features of the model are illustrated by means of graphs

Figure 10 shows, for different values of r , the functional relation between the total response time of a customer and the length of his service time. As r increases, waiting of short demands decreases while waiting of long demands increases In these examples $\theta_i = \theta$, $i=1,2,\ldots,r$, hence the unconditional expected total response time is independent of r Obviously, any decrease in waiting time of short demands is obtained at the expense of long demands. The terms "short demand" and "long demand" are used here qualitatively; in a real situation, it suffices to define an overall objective function and optimize with respect to the control parameters r and θ.

Another way of increasing priority to "short" customers is by increasing the magnitude of θ (assuming $\theta_i = \theta$, $i=1,2,\ldots,r$). This is illustrated in Fig 11, where r is kept constant for different values of θ . Note that when θ exceeds a certain magnitude, priority to short services is reduced and the queue discipline tends to FIFO. The time losses, represented by set-ups, decrease when θ is increased since the value of τ (assuming $\tau_i = \tau$, $i=1,2,\ldots,r$) is kept constant

In the preceding examples identical values of θ_i have been used in all queues, i.e., $\theta_i = \theta$, $i=1,2,\ldots,r$. The priority given to short demands may be increased, at the expense of long demands, by using a decreasing sequence $(\theta_i,\ i=1,2,\ldots,r)$ of quantum processing times. Here a R.R.r system with a fixed quantum processing time (equaling θ in all queues) is compared with an equivalent system in which θ_i is

an arbitrary sequence. It is assumed that set-up times are the same constant for all queues in both systems, i.e., $\tau_i = \tau$, $i=1,2,\ldots,r$, and furthermore that all other parameters, except for the quantum processing times, are identical in both systems. For the comparison to be meaningful, the expected time losses (represented by set-ups) per service must be equal in both systems. This implies that the expected number of quanta per service is the same in both. Fig. 12 shows an example where the sequence $(\theta_i,\ i=1,2,\ldots,50)$ is of the form $\theta_{i+1} = 0.95\,\theta_i$, $i=1,2,\ldots,50$ and $\theta_1 = 10$. The equivalent value of θ for the system where $\theta_i = \theta$, $i=1,2,\ldots,50$, is $\theta = 2.81$.

As mentioned earlier, during peak-hours (oversaturation) the R.R.r system still satisfies short demand while service of long demand is automatically postponed to non-peak hours. This is shown in Fig. 13. Table 1 supplements Fig. 13 by giving, for each traffic intensity, the longest demand that can be still served in a finite time.

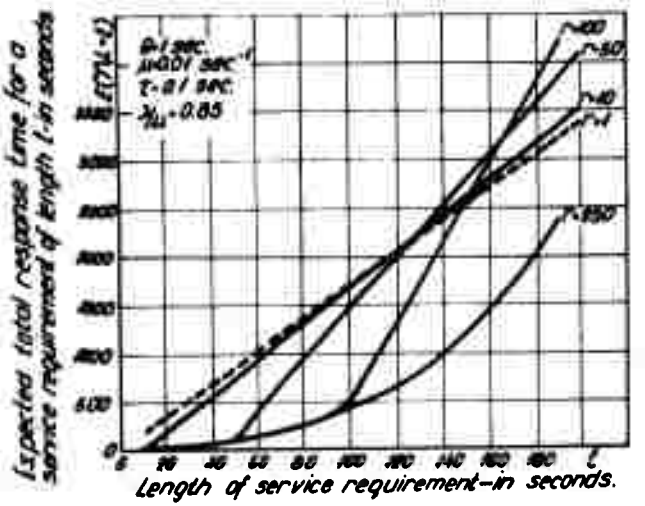

Figure 10: Expected Total Response Time for a Service Requirement of Length t as a Function of t. (for different values of r).

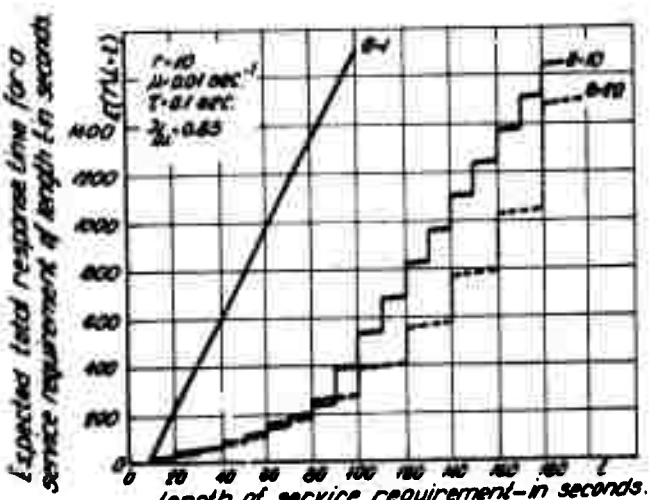

Figure 11: Expected Total Response Time for a Service Requirement of Length t as a Function of t. (For different values of θ).

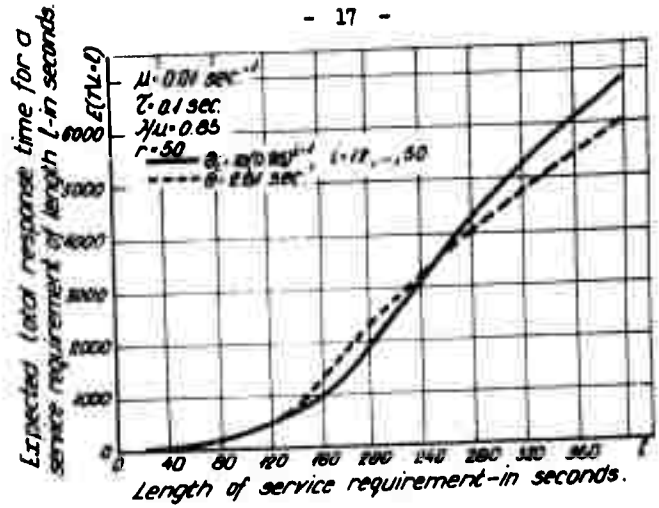

Figure 12: Expected Total Response Time for a Service Requirement of Length l as a Function of l.

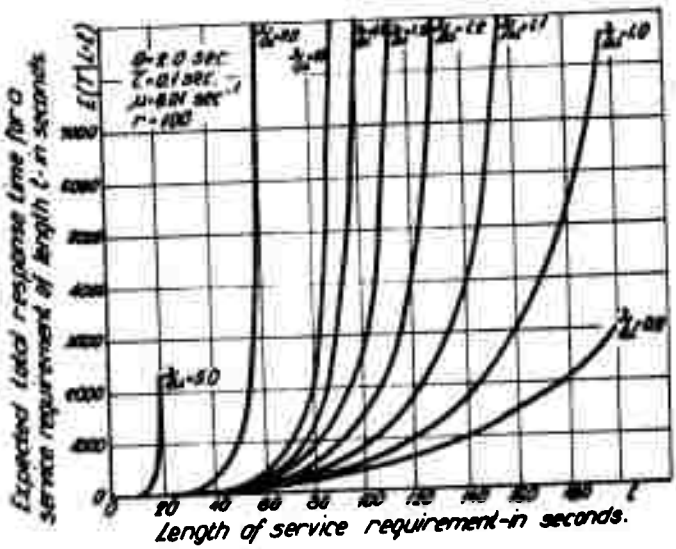

Figure 13: Expected Total Response Time for a Service Requirement of Length l as a Function of l.
(For different values of trafic intensity.)

Table 1: Largest Satisfied Demand as a Function of Traffic Intensity.
($r = 100$, $\mu = 0.01\text{sec}^{-1}$, $\tau_i = 0.1\text{sec}$, $\theta_i = 2$ sec, $i=1,2,\ldots,r$)

λ/μ	$\rho = \lambda E(S)$ *	Number of unsaturated queues	Largest satisfied demand in seconds
0.9	0.95	100	all demands
1.0	1.05	99	198
1.1	1.16	99	198
1.2	1.26	78	156
1.3	1.37	65	130
1.4	1.47	56	112
1.5	1.58	50	100
2.0	2.10	32	64
5.0	5.25	10	20

* Note that the steady state condition is $\rho < 1$.

Model V - R.R.1 With Priority Classes.

If, as is commonly the case, a single computer center dispenses service to different classes of customers and it is desired to provide better service to some of them at the expense of others, the concept of priority is introduced. In a combination of the two disciplines (i.e., a priority discipline with R.R. schedule in every class), improved service to higher priority customers is coupled with one to customers with short demands in every priority class. In a computer system, possible combinations of R.R. schedule with a priority discipline are as follows:

(1) Uninterrupted service quantum - at the end of a service quantum the next customer admitted to service is the highest priority customer present in the system.

(2) Interrupted service quantum - at any moment of time, the service station is occupied by the highest priority customer present. Here two cases are distinguished :

(a) Round-robin schedule with preemptive-repeat priority regime, where the whole interrupted quantum is lost.

(b) Round-robin schedule with mixed preemptive priority regime. Assuming an interruption in the set-up time, the given set-up is lost; but in the event of an interruption in the quantum processing time, service is resumed, after a new set-up time, at the point of interruption. This model is somewhat more realistic, since service interruption by a higher priority customer means replacement of the lower priority customer's program by the higher priority one. When there are no higher priority customers in the system, the interrupted customer's program is returned to the magnetic core memory, and the service resumed at the point it was stopped. As mentioned above, these activities are referred to as set-up activities.

The above cases were mathematically investigated by Adiri[10]. Another related situation investigated by him, of a rather theoretical nature[11], is the round-robin coupled with preemptive-resume priorities where no time losses due to interruptions are involved.

REFERENCES

Takács, L. (1963), A Single Server Queue with Feedback. The Bell System Technical Journal, pp. 505-519.

Little, J.D.C. (1961), A Proof of the Queueing Formula $L = \lambda W$. Ops.Res.,Vol.9,No.3,pp.383-387.

Adiri, I. and B. Avi-Itzhak, A Time-Sharing Queue. Mimeograph Series No. 5 of Ops. Res., Stat. and Econ., Fac. of Ind. and Mgt. Eng., Technion. (Submitted for publication).

Schrage, L.E. (1966) Some Queueing Models for Time-Shared Facility. Ph.D. thesis, Cornell University.

Adiri, I. and B. Avi-Itzhak, A Time-Sharing Queue with a Finite Number of Customers. Mimeograph Series No. 10 of Ops. Res., Stat. and Econ., Fac. of Ind. and Mgt. Eng., Technion. (Accepted for publication in the Journal of the ACM).

Takács, L. (1957) On a Stochastic Process Concerning Some Waiting Time Problems. Theory of Probability and its Applications. Vol. 11, No. 1, pp. 90-103.

Krishnamoorthi, B. and R.G. Wood, (1966) Time-Shared Operations with Both Interarrival and Service Time Exponential. Journal of the ACM, Vol. 13, No. 3, pp. 317-338.

Adiri, I. and B. Avi-Itzhak, A Time-Sharing Model with Many Queues. Mimeograph Series No. 20 of Ops. Res., Stat. and Econ., Fac. of Ind. and Mgt. Eng., Technion. (Submitted for publication).

Schrage, L.E. (1967) The Queue M/G/1 with Feedback to Lower Priority Queues. Mgt. Sci., Vol. 13, No. 7, pp. 466-474.

Adiri, I. Computer Time-Sharing Queues with Priorities. Mimeograph Series No. 13 of Ops. Res., Stat. and Econ., Fac. of Ind. and Mgt. Eng., Technion. (Submitted for publication).

Adiri, I. A Time-Sharing Queue with Preemptive-Resume Priority Discipline. Mimeograph Series No. 7 of Ops. Res., Stat. and Econ., Fac. of Ind. and Mgt. Eng., Technion. (Accepted for publication in The Israel Journal of Technology).

O.R. Mimeograph Series Publications

1.	Mitrani, I. & Avi-Itzhak, B.	"A Many Server System with Service Interruptions".
2.	Haitovsky, Y.	"Missing Data in Regression Analysis".
3.	Haitovsky, Y.	"Efficient least squares regression of differently grouped observations when the cross classifications are unknown".
4.	Powell, B.A. & Avi-Itzhak, B.	"Queueing Systems with Enforced Idle Time".
5.	Adiri, I. & Avi-Itzhak, B.	"A Time-Sharing Queue".
6.	Levin, O.	"Optimal Control of a Storage Reservoir During a Flood Season".
7.	Adiri, I.	"A Time-Sharing Queue with Preemptive Resume Priority Discipline".
8.	Haitovsky, Y.	"The Theory of Regression Estimation from Grouped Observations".
9.	Pollatschek, M.A. & Avi-Itzhak, B.	"Algorithms for Stochastic Games with Geometrical Interpretation".
10.	Adiri, I. & Avi-Itzhak, B.	"A Time-Sharing Queue with a Finite Number of Customers".
11.	Haitovsky, Y.	"A Note on the Maximization of $\bar{R}^2$".
12.	Naor, P.	"On the Regulation of Queue Size by Levying Tolls".
13.	Adiri, I.	"Computer Time-Sharing Queue with Priorities".
14.	Avi-Itzhak, B. & Mandelbaum, M.	"Mathematical Models for Outgoing Traffic Flow in an Airport Termin
15.	Avi-Itzhak, B. & Mandelbaum, M.	"Introduction to Queueing with Splitting and Matching".
16.	Pollatschek, M.A. & Levin, O.	"On the Optimization Model of Groundwater Utilization".
17.	Kondor, Y.	"Linear Programming of Income Tax Rates".
18.	mer, P.L.	"Increasing the Capacity of a Network".
19.	Hammer, P.L.	"Time Minimizing Transportation Problems".
20.	Adiri, I. & Avi-Itzhak, B.	"A Time-Sharing Model with Many Queues".

21.	Haitovsky, Y.	"Estimation of Regression Equations when a Block of Observations is Missing".
22.	Haitovsky, Y.	"Multicollinearity in Regression Analysis".
23.	Haitovsky, Y.	"Approximation Formulae for Covariances and Correlations of Products of Random Variables".
24.	Almogi, Y. & Levin, O.	"Optimal Cargo Shipping Problem with Hyperbolic Objective Function".
25.	Passy, U.	"Mass Action and Polynomial Optimization".
26.	Passy, U.	"Modular Design - An Exercise in Structured Geometric Programming".
27.	Adiri, I. & Avi-Itzhak, B.	"Queueing Models for Time-Sharing Service Systems".