UNIVERSITY OF NORTH CAROLINA
Department of Statistics
Chapel Hill, N. C.

ON DISCRETIONARY PRIORITY QUEUEING

by

B. Avi-Itzhak, I. Brosh
Technion, Israel Institute of Technology, Haifa, Israel

and

P. Naor
Technion, Israel Institute of Technology, Haifa, Israel

and

University of North Carolina, Chapel Hill, N. C.

November, 1962

This research was supported by the Office of Naval Research under contract No. Nonr-855(09) for research in probability and statistics at the University of North Carolina, Chapel Hill, N. C. Reproduction in whole or in part is permitted for any purpose of the United States Government.

Institute of Statistics
Mimeo Series No. 338

ON DISCRETIONARY PRIORITY QUEUEING[1]

B. Avi-Itzhak, I. Brosh

Technion, Israel Institute of Technology, Haifa, Israel

and

P. Naor

Technion, Israel Institute of Technology, Haifa, Israel

and

University of North Carolina, Chapel Hill, N. C.

ABSTRACT

A priority regime is envisaged for single server queueing systems composed of two customer populations with Poisson arrivals which is intermediate between the two extreme doctrines: a) head of the line priority, b) pre-emptive priority. The state of intermediacy is represented by discretionary powers vested in the server to interrupt recently initiated - and not to interrupt almost completed - service to a low priority customer upon the arrival of a high priority customer. For the case of constant service times the discretionary rule is defined and the ensuing queueing characteristics analyzed; in particular, the average total queue lengths of both high and low priority customers are derived for two different cases: a) the resume situation where service renewed to a low priority customer starts at the point of interruption; b) the repeat situation where service given to a low priority customer before an interruption, is completely lost. Optimisation procedures are outlined and for the resume situation a simple optimal discretionary rule is obtained.

---

[1]This research was supported by the Office of Naval Research under contract No. Nonr-855(09) for research in probability and statistics at the University of North Carolina, Chapel Hill, N. C. Reproduction in whole or in part is permitted for any purpose of the United States Government.

ON DISCRETIONARY PRIORITY QUEUEING

B. Avi-Itzhak, I. Brosh

Technion, Israel Institute of Technology, Haifa, Israel

and

P. Naor

Technion, Israel Institute of Technology, Haifa, Israel

and

University of North Carolina, Chapel Hill, N. C.

I. Introduction

In the literature on queueing theory mention is made of two types of priorities: a) head of the line priority, and b) pre-emptive priority. The rule pertaining to type a) is that the service rendered to a low priority* customer is never interrupted if a high priority* customer arrives; rather the HP customer is placed at the head of the waiting line and is dealt with only after completion of service to the LP customer at the station. In other words, the newly arriving HP customer does not take precedence over the LP customer in service. The rule pertaining to type b) is that a newly arriving HP customer always displaces an LP customer from the station.

Now in many practical situations in the business, industrial and military fields neither of these two extreme rules is called for, even though some priority regime must prevail. Intuitively, one is inclined to prefer the type a) rule at times when the service to the LP customer is almost completed; if, on the other hand, service to the LP customer has hardly started application of the type b) rule seems to be the proper course of action.

---

*In this study "low priority" and "high priority" will be shortened to LP and HP, respectively.

Thus it appears desirable to leave some discretion to the server (or to whoever is in charge of the system) as to the administration of the priority regime. To exercise his discretion the server must possess some further knowledge regarding the character of the serving process, actual past performance on the customer in service and future expectation of the service duration. Furthermore, the various cost components should be known to him so that the cost implications of each of the possible decisions may be evaluated.

The purpose of this study is to analyze discretionary powers with which we endow the server for the particular case of constant service times (for each priority class). A discretionary rule is easily defined and simply parametrized in such priority queueing models. The rule defined will be applied to two different situations. One of them is characterized by the (more common but frequently less realistic) assumption that interrupted service may be resumed at convenience without any loss. In the second situation it is assumed that a customer, whose service has been interrupted and later renewed, starts all over again, i.e., service must be repeated. The analysis of the resume situation leads - on introtroducing reasonable costs - to a simple optimal rule regarding the use of the discretionary powers vested in the server. The repeat situation is much more complex and no simple optimisation rule can be provided. However, the analysis of the repeat situation is carried to a point where knowledge of the numerical values of the operative parameters and of the costs enables the attainment of the optimum.

II. Model A: Two Priority Classes, Constant Service Times, Resume Situation

Consider a service station rendering service to two Poisson streams (HP and LP) of incoming customers with parameters λ_1 and λ_2, respectively. Service times are constant within each class of customers and their designation is S_1 and

S_2, respectively. Now a customer of class 1(HP) takes precedence over one of class 2 (LP); however an LP customer in service is not necessarily displaced by an incoming HP customer. Rather the server uses his discretion in the following manner: If the service time already devoted to the LP customer at the station is equal to or exceeds a value of ϕS_2, where ϕ is an arbitrary, discretionary constant $(0 \leq \phi \leq 1)$, the HP customer is placed at the head of the waiting line; if, on the other hand, the partial service time already elapsed falls short of ϕS_2 the LP customer is replaced by the incoming HP customer and moves to the head of the waiting line (of his own class). On returning later to the station his service is resumed at the point of interruption. To sum up, during the first stage (of duration ϕS_2) of service to an LP customer the pre-emptive priority rule is followed; during the second stage (of duration $(1 - \phi)S_2$) of service to an LP customer the head of the line priority rule is followed.

Now the fraction of time b_1, during which the station is busy giving service to HP customers is clearly equal to

$$b_1 = \lambda_1 S_1 \tag{1}$$

and the LP - busy fraction is given by

$$b_2 = \lambda_2 S_2 \tag{2}$$

Without any formal proof we equate the existence of steady state conditions with non-saturation the criterion of which is obviously

$$1 - b_1 - b_2 > 0 \tag{3}$$

Let the number of HP customers in the line waiting for initiation of service be denoted by w_1; the expectation of the waiting time θ_{1w}, (up to initiation of service) of a customer is connected with the expected value of w_1 by

$$E(w_1) = \lambda_1 E(\Theta_{1w}) \tag{4}$$

A further relation between $E(\Theta_{1w})$ and $E(w_1)$ may be obtained by the following line of reasoning. An incoming HP customer will encounter an average waiting line of length $E(w_1)$; this will contribute $E(w_1)S_1$ to his average waiting time. In addition, with probability b_1, he will find an HP customer being serviced and, with probability $(1 - \phi)b_2$, he encounters an LP customer in the second stage of his service. The average contributions of these two events if they occur, to the average waiting time are $S_1/2$ and $(1 - \phi)S_2/2$, respectively.

Thus we establish

$$E(\Theta_{1w}) = E(w_1)S_1 + b_1 \frac{S_1}{2} + (1 - \phi)b_2 (1 - \phi) \frac{S_2}{2} \tag{5}$$

Combination of (4) and (5) yields

$$E(w_1) = \frac{b_1^2 + \lambda_1 S_2 b_2 (1 - \phi)^2}{2(1 - b_1)} \tag{6}$$

The total expected queue, q_1, of HP customers at the station equals

$$q_1 = b_1 + E(w_1) = b_1 + \frac{b_1^2 + \lambda_1 S_2 b_2 (1 - \phi)^2}{2(1 - b_1)} \tag{7}$$

Choosing $\phi = 1$ is equivalent to the establishment of a pre-emptive priority regime and the HP customers' average queue length should be equal to that of a simple queueing model with Poisson input and constant service times. This is indeed the case.

Selecting $\phi = 0$ is equivalent to the establishment of a head of the line priority regime. The average queue length reduces to

$$q_1(\phi = 0) = b_1 + \frac{b_1^2 + \lambda_1 S_2 b_1}{2(1 - b_1)} \tag{8}$$

and this may also be obtained from Cobham's (1954) study who deals with arbitrary distributions of service times in a head of the line priority regime.

For the evaluation of the average queue length of LP customers we note, first, that a relation parallel to (4) holds

$$E(w_2) = \lambda_2 E(\theta_{2w}) \tag{9}$$

where w_2 and θ_{2w} represent the waiting line of LP customers and the waiting time of a single LP customer, respectively.

Next we observe that $E(\theta_{2w})$ is invariant under the various priority rules. This must be the case since LP customers whose service has not started yet - and only these are defined as waiting - are completely unaffected by changing the regime from one extreme through states of intermediacy ($\phi \neq 0, 1$) to the other extreme. Indeed Cobham (1954) and Miller (1960) were concerned with the cases $\phi = 0$ and $\phi = 1$, respectively, and both of them obtained the first moment* as

$$E(\theta_{2w}) = \frac{\lambda_1 s_1^2 + \lambda_2 s_2^2}{2(1-b_1)(1-b_1-b_2)} \tag{10}$$

Obviously (10) must hold also in the intermediate model under consideration. Combination of (9) and (10) generates the average waiting line

$$E(w_2) = \frac{\lambda_2 \lambda_1 s_1^2 + \lambda_2 s_2^2}{2(1-b_1)(1-b_1-b_2)} = \frac{b_1 \lambda_2 s_1 + b_2^2}{2(1-b_1)(1-b_1-b_2)} \tag{11}$$

Consider now the residence time, T, of an LP customer at the station; this is defined as the time which elapses from initiation to termination of service. The residence time is made up of the time devoted to the service of the

---

*Miller (1960) also obtained higher moments.

LP customer himself and of the service times of HP customers who arrive during the time at which it is possible to displace the LP customer or to keep him in his displaced position. Now whenever an LP customer is displaced from the station a period t_1 elapses until service to him is renewed. Miller (1960) and Avi-Itzhak and Naor (1962) obtained the expected value of t_1 as

$$E(t_1) = \frac{S_1}{1-b_1} \tag{12}$$

If an LP customer is displaced k times from the service station the expected value of his residence time equals

$$E(T|k) = S_2 + k\, E(t_1) = S_2 + \frac{k\, S_1}{1-b_1} \tag{13}$$

The number of displacements, k, is, of course, a Poisson-distributed random variable with parameter $\lambda_1 \phi S_2$

$$p(k) = e^{-\lambda_1 \phi S_2} \frac{(\lambda_1 \phi S_2)^k}{k!} \tag{14}$$

The unconditional expected residence time is then equal to

$$E(T) = \sum_{k=o}^{\infty} p(k)\, E(T|k) = S_2 \sum_{k=o}^{\infty} p(k) + \frac{S_1}{1-b_1} \sum_{k=o}^{\infty} k\, p(k) = S_2\left(1 + \frac{\phi\, b_1}{1-b_1}\right) \tag{15}$$

The average queue length of LP customers, q_2, is now obtained as

$$q_2 = \lambda_2\, E(T) + E(w_2) = b_2\left(1 + \frac{\phi\, b_1}{1-b_1}\right) + \frac{b_1 \lambda_2 S_1 + b_2^2}{2(1-b_1)(1-b_1-b_2)} \tag{16}$$

Again, setting $\phi = 0$ and $\phi = 1$, respectively, leads to well-known results relating to the two extreme priority regimes.

We can now turn to economic considerations and arrive - on making cost assumptions - at an optimal value of ϕ, ϕ^* say. One reasonable cost structure only will be dealt with here: it will be assumed that the cost of delaying a

single HP customer in the queue equals c_1 per unit time; and the cost of delaying a single LP customer is taken to be c_2 per unit time. Thus the average total cost, c, per unit time is equal to

$$c = c_1 q_1 + c_2 q_2 = c_1 b_1 + \frac{c_1 b_1^2 + c_1 \lambda_1 S_2 b_2 (1-\phi)^2}{2(1 - b_1)} + c_2 b_2 \left(1 + \frac{\phi\, b_1}{1-b_1}\right) + \\ + \frac{c_2 b_1 \lambda_2 S_1 + c_2 b_2^2}{2(1-b_1)(1-b_1-b_2)} \tag{17}$$

On differentiating and setting the derivative equal to zero ϕ^* is evaluated as

$$\phi^* = 1 - \frac{c_2 S_1}{c_1 S_2} \tag{18}$$

This result is meaningful if, and only if, the following inequality holds

$$c_2 S_1 \leq c_1 S_2 \tag{19}$$

If inequality (19) is not realized it may be shown by elementary but lengthy considerations that the optimum is reached if class 1 and class 2 customers exchange their priority categories. The optimum value of ϕ is still given by (18) with indices interchanged.

It is interesting to note that ϕ^* is independent of the arrival rates, λ_1 and λ_2.

III. Model B: Two Priority Classes, Constant Service Times, Repeat Situation

The priority queueing model with which we deal in this Section is identical with Model A in all but one respect. In the following it will be assumed that interrupted service is virtually lost. Whenever an LP customer is displaced

in the midst of his service he will - on renewal of service - require the attention of the station during an uninterrupted period of duration S_2; prior service must be repeated.

We observe that the average queue length (and, indeed, all other properties) of the HP customers is the same in both resume and repeat situations. Thus relation (7) holds in Model B as well.

To evaluate properties of the queue of LP customers the concept of gross service time, S_g, must be developed.

An LP customer is served by the station and his service is subjected to $k(=0, 1, \ldots)$ interruptions by HP customers where k is a geometrically distributed random variable

$$p(k) = e^{-\lambda_1 \phi S_2}(1 - e^{-\lambda_1 \phi S_2})^k \tag{20}$$

the expected value of which equals

$$E(k) = \sum_{k=0}^{\infty} k\, p(k) = e^{\lambda_1 \phi S_2} - 1 \tag{21}$$

Now whenever displacement has taken place the partial service time $\tau(0 \leq \tau \leq \phi S_2)$ expended is lost. This partial service time, τ, is a random variable with a density

$$h(\tau) = \frac{\lambda_1 e^{-\lambda_1 \tau}}{1 - e^{-\lambda_1 \phi S_2}} \tag{22}$$

and an expectation

$$E(\tau) = \int_0^{\phi S_2} \tau\, h(\tau)\, d\tau = \frac{1}{\lambda_1} - \frac{\phi S_2}{e^{\lambda_1 \phi S_2} - 1} \tag{23}$$

An LP customer occupies the service station during a period of duration

$$S_g = S_2 + \sum_{i=1}^{k} \tau_i \qquad (24)$$

where both k and τ_i are random variables; the expected value of S_g is obtained as

$$E(S_g) = \sum_{k=o}^{\infty} p(k)\, E(S_g|k) = \sum_{k=o}^{\infty} p(k) \left[S_2 + k\, E(\tau) \right] =$$

$$= S_2 + E(k)\, E(\tau) = (1-\phi)S_2 + \frac{e^{\lambda_1 \phi\, S_2} - 1}{\lambda_1} = (1-\phi)S_2 + \frac{E(k)}{\lambda_1} \qquad (25)$$

This is the expected value of the gross service time.

The fraction of time, b_2, during which the station is engaged in giving service to LP customers equals

$$b_2 = \lambda_2\, E(S_g) = \lambda_2 S_2 (1 - \phi) + \frac{\lambda_2}{\lambda_1} \left(e^{\lambda_1 \phi\, S_2} - 1 \right) \qquad (26)$$

The busy fraction, b_1, relating to HP customers is given, of course, by relation (1). Non-saturation is again assumed to be equivalent to the existence of steady state conditions; the criterion continues to be inequality (3) with b_2 redefined by (26).

To gain further physical insight into the model we wish to determine the average residence time, $E(T)$, of an LP customer at the station and the average time interval, $E(U)$, which elapses from the initiation of service to an LP customer to that epoch at which the station is prepared to give service (if requested) to the next LP customer.

Consider an LP customer whose service is interrupted (and repeated) k times. The arrival of (and interruption by) an HP customer causes the station to be engaged to HP customers during an average time interval $E(t_1)$ given by expression (12). Clearly the expected residence time of the LP customer who is

interrupted k times equals

$$E(T|k) = E(S_g|k) + k\,E(t_1) \tag{27}$$

and the expected unconditional residence time is obtained as

$$E(T) = \sum_{k=o}^{\infty} p(k)\,E(T|k) = \sum_{k=o}^{\infty} p(k)\,\left[E(S_g|k) + k\,E(t_1)\right] =$$

$$= E(S_g) + E(k)\,E(t_1) = (1-\phi)S_2 + \frac{E(k)}{\lambda_1} + \frac{E(k)S_1}{1-b_1} \tag{28}$$

Further rearrangement yields

$$E(T) = \frac{E(k)}{\lambda_1(1-b_1)} + (1-\phi)S_2 \tag{29}$$

The first term on the right hand side of (29) measures the average time which elapses from the initiation of service to an LP customer up to the time at which he overcomes his inferior standing by completing the part ϕS_2 of his service. The second term represents the remaining part of his service during which it is no longer possible to displace him. At the beginning of this latter part no HP customers are present at the station. Hence, on the average, the number of HP customers in the waiting line at the time of termination of service to the LP customer equals

$$E(w_1|\text{ at termination of service to LP customer}) = \lambda_1(1-\phi)S_2 \tag{30}$$

As a single new HP customer gives rise to an HP-busy period of average duration $E(t_1)$, completion of service to an LP customer will be followed by an HP-busy period the average length of which is given by the product of the right hand sides of (12) and (30). For the evaluation of $E(U)$ this must be added to the average residence time.

$$E(U) = E(T) + \frac{\lambda_1(1-\phi)S_2S_1}{1 - b_1} = \frac{E(k)}{\lambda_1(1-b_1)} + (1 - \phi)S_2 + \frac{b_1(1-\phi)S_2}{1 - b_1} =$$

$$= \frac{E(k)}{\lambda_1(1-b_1)} + \frac{(1-\phi)S_2}{1-b_1} = \frac{E(k) + \lambda_1(1-\phi)S_2}{\lambda_1(1-b_1)} = \frac{E(S_g)}{1-b_1} \qquad (31)$$

An alternative definition of non-saturation (and the existence of steady state conditions) is then given by

$$\lambda_2 E(U) < 1 \qquad (32)$$

It is not difficult to see that (3) and (32) are equivalent.

To evaluate the average queue of LP customers we proceed by considering the possible states of the system in some more detailed fashion. In the following we shall distinguish between unprimed states, E_{mn}, and primed states, E'_{mn}. Whenever the system is in an unprimed state E_{mn} $(0 \leq m < \infty,\ 0 \leq n < \infty)$ there are m HP customers and n LP customers in the queue; if $m > 0$ an HP customer is being served, and if $m = 0$ (and $n > 0$) an LP customer is being served but the future service required of the station exceeds $(1-\phi)S_2$. Whenever the system is in a primed state E'_{mn} $(0 \leq m < \infty,\ 1 \leq n < \infty)$ there are again m HP customers and n LP customers in the queue but the customer in service is of the LP variety.

In other words, a state is considered unprimed if the shortest future service time of an LP customer exceeds $(1-\phi)S_2$; it is considered primed if this shortest future service time of an LP customer falls short of (or is equal to) $(1-\phi)S_2$.

The stationary probabilities pertaining to these events will be designated as p_{mn} and p'_{mn}, respectively. Clearly the following must hold

$$\sum_{m=0}^{\infty}\sum_{n=0}^{\infty} p_{mn} + \sum_{m=0}^{\infty}\sum_{n=1}^{\infty} p'_{mn} = 1 \tag{33}$$

It is useful to introduce further definitions

$$p_{m.} = \sum_{n=0}^{\infty} p_{mn} \tag{34}$$

$$p_{.n} = \sum_{m=0}^{\infty} p_{mn} \tag{35}$$

$$p'_{m.} = \sum_{n=1}^{\infty} p'_{mn} \tag{36}$$

$$p'_{.n} = \sum_{m=0}^{\infty} p'_{mn} \tag{37}$$

The fraction of time during which the station is HP-busy, b_1, must equal

$$b_1 = \lambda_1 S_1 = \sum_{m=1}^{\infty} p_{m.} \tag{38}$$

It may be shown in a straightforward manner that

$$\frac{\lambda_2}{\lambda_1} E(k) = \sum_{n=1}^{\infty} p_{on} = p_{o.} - p_{oo} \tag{39}$$

and

$$\lambda_2(1 - \phi)S_2 = \sum_{m=1}^{\infty} p'_{.n} \tag{40}$$

The fraction of time during which the station is LP-busy, b_2, is equal to the sum of (39) and (40)

$$b_2 = \frac{\lambda_2}{\lambda_1} E(k) + \lambda_2 S_2(1 - \phi) = \sum_{n=1}^{\infty} p_{on} + \sum_{n=1}^{\infty} p'_{.n} \tag{41}$$

In principle, the average queue length of LP customers obtained is given by

$$q_2 = E(n) = \sum_{n=1}^{\infty} n\,(p_{.n} + p'_{.n}) \tag{42}$$

In practice we proceed as follows: Consider an LP customer who arrives at the service station while service is given to an HP customer; in other words his arrival occurs when the system is in an unprimed state (excluding E_{on}). His (average) total queueing time is made up of three components:

a) On the average the time he waits for termination of service to the HP customers who are physically present at the time of his arrival equals $\frac{1}{2} S_1$ (for the HP customer in service) plus $S_1(\sum_{m=1}^{\infty} p_{m.})^{-1} \sum_{m=1}^{\infty} (m-1)p_{m.}$ (for the HP customers in the waiting line). However, by an argumentation similar to that leading to (12) and (31) this average is increased by a factor of $(1-b_1)^{-1}$; this factor is introduced to take into account that further HP customers arrive while service is rendered to HP customers present at the time of the LP customer's arrival. Thus the contribution of this term to average total queueing time of an LP customer equals

$$S_1(1-b_1)^{-1} \left[\frac{1}{2} + (\sum_{m=1}^{\infty} p_{m.})^{-1} \sum_{m=1}^{\infty} (m-1)\, p_{m.} \right] .$$

b) Each LP customer who is in the queue at the time of the new LP customer's arrival delays service for an average period $E(U)$; this quantity was evaluated in equation (31). The average number of LP customers present in the queue at that time equals $\sum_{m=1}^{\infty} \sum_{n=0}^{\infty} n\, p_{mn}$. Thus the contribution of this term to the average total queueing time equals $E(U) \sum_{m=1}^{\infty} \sum_{n=0}^{\infty} n\, p_{mn} (\sum_{m=1}^{\infty} p_{m.})^{-1}$.

c) The third term contributing to the average total queueing time is the residence time, $E(T)$, of the new LP customer.

Thus we obtain the average of the total queueing time of an LP customer who arrives while service is rendered to an HP customer as

$$E(\theta_{2q}\,|\,{}_{\substack{m>0\\ n\geq 0}}\,E_{mn}) = \frac{S_1}{2(1-b_1)} + \frac{S_1}{1-b_1}\,\frac{\sum_{m=1}^{\infty}(m-1)p_{m.}}{\sum_{m=1}^{\infty}p_{m.}} +$$

$$+\,\frac{E(U)\sum_{m=1}^{\infty}\sum_{n=0}^{\infty}n\,p_{mn}}{\sum_{m=1}^{\infty}p_{m.}} + E(T) \qquad (43)$$

Now if an LP customer arrives while the system is in a primed state (that is, service in its terminal stage is rendered to an LP customer) a relation parallel to (43) may be set up by a similar line of argumentation

$$E(\theta_{2q}\,|\,{}_{\substack{m>0\\ n\geq 0}}\,E'_{mn}) = \frac{(1-\phi)S_2}{2(1-b_1)} + \frac{S_1}{1-b_1}\,\frac{\sum_{m=0}^{\infty}m\,p'_{m.}}{\sum_{m=0}^{\infty}p'_{m}} +$$

$$+\,\frac{E(U)\sum_{m=0}^{\infty}\sum_{n=1}^{\infty}(n-1)p'_{mn}}{\sum_{n=1}^{\infty}p'_{.n}} + E(T) \qquad (44)$$

An LP customer who arrives when the system is in state E_{oo} goes immediately into "residence". Thus his average total queueing time equals

$$E(\theta_{2q}|E_{oo}) = E(T) \qquad (45)$$

The last case to be considered is the encounter of the newly arriving LP customer with a state $E_{on}(n > 0)$. Such a state is characterized by the possibility that the LP customer in service will be displaced from the service station; this contingency will occur if an HP customer arrives before completion of partial service time ϕS_2. Consider, then, the two subcases: a) the LP customer in service is able to complete it; b) the LP customer in service is

displaced by an incoming HP customer.

The probability, π, of subcase a) to occur may be obtained by observing that the average time the system spends in $\bigcup_{n>0} E_{on}$ within the average gross service time to an LP customer equals $E(k)\,E(\tau) + \phi\, S_2$. The first of these two terms indicates, of course, the occasions on which service to the LP customer is interrupted; the second term represents the successful "breakthrough" of the LP customer towards unhampered completion of service. Hence the probability π equals

$$\pi = \frac{\phi\, S_2}{\phi\, S_2 + E(k)E(\tau)} = \frac{\lambda_1\, \phi\, S_2}{E(k)} \tag{46}$$

The average total queueing time for subcase a) is derived in the usual fashion

$$E\left(\theta_{2q} \Big| \bigcup_{n>0} E_{on},\ \text{subcase a)}\right) = \frac{\phi\, S_2}{2} + \frac{(1-\phi)S_2}{1-b_1} + \frac{E(U)\sum_{n=1}^{\infty}(n-1)\,p_{on}}{\sum_{n=1}^{\infty} p_{on}} + E(T) \tag{47}$$

In subcase b) we note that the new LP customer arrives during a τ-interval and the "future tail", τ^*, of this interval possesses a density which is the random modification of the original density $k(\tau)$. The quantity of interest is the expected value of τ^* and for its derivation the second moment of the original density is required. On making use of (22) we obtain

$$E(\tau^2) = \int_0^{\phi S_2} \tau^2\, h(\tau)\, d\tau = \frac{2}{\lambda_1^2} - \frac{\phi\, S_2\left(\frac{2}{\lambda_1} - \phi\, S_2\right)}{E(k)} \tag{48}$$

The expected value of τ^* is then equal to

$$E(\tau^*) = \frac{E(\tau^2)}{2E(\tau)} = \frac{\frac{2}{\lambda_1^2} - \frac{\phi S_2(\frac{2}{\lambda_1} - \phi S_2)}{E(k)}}{\frac{2}{\lambda_1} - \frac{2\phi S_2}{E(k)}} \tag{49}$$

The average total queueing time is derived

$$E(\theta_{2q} | \bigcup_{n>0} E_{on}, \text{ subcase b)}) = E(\tau^*) + \frac{S_1}{1-b_1} + \frac{E(U) \sum_{n=1}^{\infty} n\, p_{on}}{\sum_{n=1}^{\infty} p_{on}} \tag{50}$$

We are now in a position to establish an equation evaluating the unconditional average total queueing time of an LP customer

$$E(\theta_{2q}) = E(\theta_{2q} | \bigcup_{\substack{m>0 \\ n\geq 0}} E_{mn}) \sum_{m=1}^{\infty} \sum_{n=o}^{\infty} p_{mn} +$$

$$+ E(\theta_{2q} | \bigcup_{\substack{m>0 \\ n\geq 0}} E'_{mn}) \sum_{m=o}^{\infty} \sum_{n=1}^{\infty} p'_{mn} + E(\theta_{2q} | E_{oo}) p_{oo} +$$

$$+ \Big[\pi E(\theta_{2q} | \bigcup_{n>0} E_{on}, \text{ subcase a}) + (1-\pi) E(\theta_{2q} | \bigcup_{n>0} E_{on}, \text{ subcase b}) \Big] \sum_{n=1}^{\infty} p_{on} \tag{51}$$

On using relations obtained earlier in this paper equation (51) is simplified to

$$E(\Theta_{2q}) = \frac{1}{1-b_1}\left[\frac{b_1}{\lambda_1}\left(q_1 - \frac{b_1}{2}\right) + \frac{b_2}{\lambda_2}(q_2 + 1) + \frac{E(k)\lambda_2}{\lambda_1}\left(\frac{1}{\lambda_1} - S_2\right) - \right.$$

$$\left. - \frac{\phi S_2 \lambda_2}{\lambda_1} - (1 - \phi)S_2\left(b_1 + \frac{\lambda_1 S_2}{2}\right)\right] \qquad (52)$$

Under conditions of statistical equilibrium the average queue and the average queueing time are connected by

$$q_2 \equiv E(n) = \lambda_2 E(\Theta_{2q}) \qquad (53)$$

Thus multiplication of (52) by λ_2 and rearrangement leads to

$$q_2 = \frac{\lambda_1\lambda_2 S_1^2 + (1-\phi)^2\lambda_2^2 S_2^2}{2(1-b_1)(1-b_1-b_2)} + \frac{b_2\frac{\lambda_1+\lambda_2}{\lambda_1} - \frac{\lambda_2^2 S_2}{\lambda_1}\left[E(k)+1\right] - (1-\phi)\lambda_2 S_2\left[b_1 + (1-\phi)\lambda_2 S_2\right]}{1 - b_1 - b_2} \qquad (54)$$

If the value $\phi = 0$ is chosen and inserted in (54) Cobham's (1954) formula for the head of the line priority case is obtained as in Section II. If the value $\phi = 1$ is selected the model reduces to a special case of pre-emptive repeat priority queueing. Indeed the relation generated from (54) is equivalent to a specialization of a result of Avi-Itzhak (1962).

The selection of an optimal ϕ - assuming that some cost structure similar to (17) properly describes the state of affairs - is not simple and no compact formula of type (18) is attainable. However, for any given set $(\lambda_1, S_1, C_1, \lambda_2, S_2, C_2)$ numerical computation of ϕ^* is not beyond the reach of a simple desk calculating machine.

IV. Conclusion

The models discussed in this paper represent special (and relatively simple) cases of a wider field designated here as "discretionary priority queueing". In actual practice a priority regime prevails in most queueing situations, and some discretion is allowed to the server (or to the agency which actually controls the situation) with respect to the exercise of the priority doctrine.

The purpose of this study was to define possible (discretionary) courses of action and to analyze the consequences of the actual choice taken. However, as said before, two specialized models only were considered and future research will have to generalize the analysis in (at least) three directions:

a) The replacement of the two customer populations assumption by a many customer populations assumption.

b) The introduction of variability into service times.

c) The exercise of the server's discretionary powers not only on the basis of actual past and expected future attention time to the customer in service but also consideration of the state of the whole system.

Furthermore, the problems may be posed when many servers are available. Also cost structures other than (17) may reveal the true state of affairs and a wide range of such alternatives may be worth investigation.

Some further study is in progress.

REFERENCES

[1] Avi-Itzhak, B. (1962). Pre-emptive Repeat Priority Queues as a Special Case of the Multi-Purpose Server Problem - II, Technion, Israel Institute of Technology. (To be published)

[2] Avi-Itzhak, B. and Naor, P. (to be published in Operations Research, Vol. 11 (1963)). Some Queueing Problems When the Service Station Is Subject to Breakdown.

[3] Cobham, A. (1954). Operations Research 2, 70-76 .

[4] Miller, R. G. (1960). Ann. Math. Stat. 31, 86-103 .