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

MULTI-PURPOSE SERVICE STATIONS IN QUEUEING PROBLEMS**

by

B. Avi-Itzhak
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.

June, 1963

*This author's research - while at the University of North Carolina - has been supported by O.N.R. Contract No. Nonr-855(09). Reproduction in whole or in part is permitted for any purpose of the United States Government.

**This paper is to be presented at the Third International Conference on Operational Research, Oslo, July, 1963.

ABSTRACT

This paper reviews a number of queueing models which deal with situations where a single service station has to concern itself with a number of missions and at any given time one of them only can be performed.

Institute of Statistics
Mimeo Series No. 363

MULTI-PURPOSE SERVICE STATIONS IN QUEUEING PROBLEMS

by

B. Avi-Itzhak

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.

= =

INTRODUCTION

The literature reporting research in queueing theory has branched out considerably during the past few years. Inter alia, problems of priority queueing as well as queueing problems with interruptions and breakdowns (random and otherwise) have received attention and thorough treatment. Problems of these types arise in situations of physically divergent character but their methodological and phenomenological similarity was recognized rather early (White and Christie (1958)). Research was undertaken in the Department of Industrial and Management Engineering, Technion, Israel Institute of Technology with the purpose a) to present a qualitatively unified treatment of the various models recognized as being representative of these situations, and b) to derive results in this area (known and as yet unknown) in a simple and direct manner. This communication will review investigations carried out over the last few years; most of the results of this research were or are being reported elsewhere in various

*This author's research - while at the University of North Carolina - has been supported by O.N.R. Contract No. Nonr-855(09). Reproduction in whole or in part is permitted for any purpose of the United States Government.

publications and one purpose of this review is to present the collection of models in some orderly fashion. An attempt will be made to link up our results with those obtained by other authors prior to, or concurrent with, our investigations.

GENERAL FEATURES OF THE MODELS AND OF THEIR ANALYSES

The queueing systems considered here have a number of features in common. In particular, it will be assumed throughout that customers arrive in stationary Poisson streams and that they are given service by a single station. From time to time this station becomes unavailable to customers and the viewpoint taken is that a station has additional missions to perform, during which no service can be rendered to customers. This is equivalent to stating that the single station possesses multi-purpose character; the various models differ in the mode of transition from mission to mission and this will be discussed in some detail for each model.

In the analysis it is always assumed (without the benefit of a rigorous mathematical proof which, though, is not too difficult to supply) that non-saturation - to be defined in each model - can be identified with (i.e. is necessary and sufficient for) the existence of stationary conditions. In the same vein, a theorem recently proved by Little (1961) in a rigorous fashion will be generalized and used throughout. An operational description of the generalized Little theorem would be the following:

In any queueing system we are concerned with random variables which are dimensionless as well as with those which possess the physical dimension of time; for each random variable with dimension "time" - describing a property of a single customer - an analogous dimensionless random variable can be found which represents the number of customers associated with this property at some instant; we have then

as a statement of the generalised Little theorem that the ratio of the expected values of these analogous quantities is equal to the reciprocal* of the customer arrival intensity, λ_1. Thus, for instance, the following relations are all manifestations of the generalised theorem

$$\frac{E(t_s)}{b} = \frac{E(t_w)}{w} = \frac{E(t_q)}{q} = \frac{1}{\lambda_1} \qquad (1)$$

where t_s, t_w, and t_q are service time, waiting time (up to initiation of service) and total queueing time, respectively, and b, w and q are the expected values of the number of customers being served**, waiting and in the total queue, respectively.

An extremely useful device (though usually insufficient by itself for the derivation of expected queue sizes, queueing times etc.) is the introduction of the cycle concept. The station is viewed as it goes through various phases - e.g. busy, idle, broken-down, setting-up etc. - and, usually, elementary considerations suffice to establish the probabilities associated with the phases; further considerations involving the notion of the busy period lead to the evaluation of the expected duration of the various phases and of the whole cycle. At that stage it is frequently possible to generate a relation - additional to those of type (1) - between expected values of dimensionless random variables and "time" random variables by following the expected life history of a newly-arrived customer at the station. This additional relation is usually of comparative simple structure if the customer originated from a stationary Poisson stream which, of course, is a basic assumption (as stated above) throughout all the models presented here. Combination of the cycle concept and the customer's expected life history ordinarily yields (by very simple algebra) the first moments of desired quantities such

*This quantity is identical with the expected interarrival time of customers.

**This quantity is identical with the busy fraction of the service station.

as queue size, queueing time, etc.

An important point to bear in mind for setting up this relation is the following: Consider a situation where the station is engaged in some activity - e.g. serving a preceding customer - at the time of arrival of a new customer whose life history at the station we propose to study. Assume that the newly arriving customer is not going to interfere with this activity; if such is the case the time interval between the arrival of the customer and the termination of the activity is the forward delay (Cox and Smith (1954), studied earlier in different contexts by Smoluchowski (1915) and Palm (1943)) of the activity duration. The density of the forward delay is the random modification (Naor (1957)) of the density of the original random variable. The expected value of the forward delay, τ say, bears a simple relationship to the first two moments of the original random variable t (activity duration), to wit

$$E(\tau) = \frac{E(t^2)}{2E(t)} = E(t)\ \frac{1+\gamma^2}{2} \qquad (2)$$

where γ is the coefficient of variation of t. If the new customer arrives "in the midst" of some activity the coefficient of variation is incorporated in the desired additional relation and makes its appearance in the final formula expressing average queue size or some such quantity in terms of the original system parameters and functional forms. Indeed the well known Khintchine - Pollaczek formula is the simplest example of such a result.

In some of the models reported it is not too difficult to evaluate higher moments of quantities under study and this is achieved either by direct means or through the setting-up of a moment generating function.

There are various other devices which are useful in some situations; they will be introduced and stated on discussing specific models.

Cost structures and optimization procedures will make their appearance only in those models where the station breakdown or the priority ordering form part of the management doctrine.

Our studies follow and generalise several aspects of work by Cobham (1954), Kesten and Runnenburg (1957), White and Christie (1958), and Miller (1960). Some of our results bear close relationship to those obtained recently by Heathcote (1961), Gaver (1962), Keilson (1962) and Thiruvengadam (1963). This review summarizes research reported in a number of papers, to wit: Avi-Itzhak and Naor (1961), Avi-Itzhak and Naor (1963), Avi-Itzhak (1963), Brosh and Naor (1963), Yadin and Naor (1963), Avi-Itzhak (1964) and Avi-Itzhak, Brosh and Naor (1964).

We shall endeavor to present the material in a unified notation. However, one exception will have to be tolerated: λ_1 denotes the stationary Poisson arrival intensity of customers in Models I to VII (as well as in Model XIV) whereas λ_2 is associated with the Poisson breakdown intensity* of the service station. In contra-distinction to the above usage in those models which are directly concerned with priority classes λ_1, λ_2, etc. will denote customer arrival intensities in this priority order. The probabilities of the station, being in working order and broken-down are designated as p_o and p_1, respectively. Service time and station repair time are represented by t_s and t_r , and their density functions by $f_s(t)$ and $f_r(t)$; ordinarily, the only assumption we make regarding them is the existence of (at least) the first two moments. Finally, in this Section, we introduce the concept of the residence time, T, of a customer: this is the time elapsing between initiation and termination of service to the customer. On the average this quantity exceeds pure service time since breakdowns may occur while service is rendered to the customer.

---

*The concept of Poisson breakdown intensity is not used in Model V. Hence no λ_2 is to be found in the equations pertaining to this model.

We note that two different regimes may exist in the queueing system. In one case, as soon as the station becomes operative after a breakdown, service to the resident customer is resumed at the point of interruption. In the other case, service rendered to the resident customer prior to the interruption is lost and must be repeated. Models with either regime will be considered in the following.

MODEL I

This simplest model of the type under study is one in which breakdowns are homogeneous in time and independent of customers' arrivals and the station's service; furthermore the existence of a resume regime is assumed.

The cycle consists of two phases - operative and broken-down - and it is easy to verify that

$$p_o = \frac{1}{1 + \lambda_2 E(t_r)} \tag{3}$$

and

$$p_1 = \frac{\lambda_2 E(t_r)}{1 + \lambda_2 E(t_r)} \tag{4}$$

Non-saturation is defined by the inequality

$$p_o - b > 0 \tag{5}$$

and this is necessary and sufficient for stationary conditions to prevail.

A little reflection leads to the density function $\phi(T)$, of the residence time

$$\phi(T) = \sum_{x=0}^{\infty} \left(\int_0^{\infty} f_s(t) \frac{(\lambda_2 t)^x}{x!} e^{-\lambda_2 t} dt \right) \left\{ f_s(t) \right\} * \left\{ f_r^{x*}(t) \right\} \tag{6}$$

where $*$ is the convolution operator and f^{x*} is the x-fold convolution of a density with itself.

From (6) the expected value of T is evaluated as

$$E(T) = \frac{E(t_s)}{p_o} \tag{7}$$

which could have been anticipated on intuitive grounds.

We have then (on the average) b/p_o customers in residence and if the average waiting line, w, is defined as consisting of those customers who wait for initiation of service we can derive its value (by means outlined in the previous Section) as

$$w = q - \frac{b}{p_o} = \frac{\lambda_1 E(t_r)(\gamma_r^2 + 1)p_o p_1 + b^2(\gamma_s^2 + 1)/p_o}{2(p_o - b)} \tag{8}$$

This is a generalisation of previously known results such as of the Khintchine - Pollaczek relation and of a formula obtained by White and Christie (1958) for exponential service and repair times.

MODEL II

In many systems the interruption mechanism is associated with the service activity and breakdowns are liable to occur only during the busy fraction of the station. The regime of service renewal is again assumed to be of the resume variety. The cycle, then, consists of two phases and their associated probabilities are given by

$$p_o = 1 - \lambda_1\lambda_2 E(t_s) E(t_r) = 1 - b\lambda_2 E(t_r) \tag{9}$$

and

$$p_1 = \lambda_1\lambda_2 E(t_s) E(t_r) = b \lambda_2 E(t_r) \tag{10}$$

The residence time density is reported by Gaver (1962) and the first two moments are

$$E(T) = E(t_s)\left[1 + \lambda_2 E(t_r)\right] \tag{11}$$

and

$$E(T^2) = E(t_s^2)\left[1 + \lambda_2 E(t_r)\right]^2 + \lambda_2 E(t_s)\, E(t_r^2) \tag{12}$$

If we employ the device of considering an interruption as an extension of service it is possible to obtain an expression for the expected waiting line almost immediately. One further definition is necessary for this purpose: Let a modified busy fraction, B, be defined as the fraction of time during which the station is either rendering service or is being repaired. This is identical for Model II with the average number of customers in residence. We have then

$$B = \lambda_1 E(T) = b\left[1 + \lambda_2 E(t_r)\right] = b + p_1 \tag{13}$$

Incidentally, by defining B we have established the criterion of non-saturation as

$$B < 1 \tag{14}$$

The expected waiting line is then given by the Khintchine - Pollaczek relation with the busy fraction replaced by the modified busy fraction and the coefficient of variation relating to residence time rather than to service time

$$w = \frac{B^2}{2(1-B)}\left(\gamma^2 + 1\right) = \lambda_1 \frac{\lambda_2 b\, E(t_r^2) + \lambda_1 \left(\frac{b + p_1}{b}\right)^2 E(t_s^2)}{2(p_o - b)} \tag{15}$$

MODEL III

Consider now a queueing system with time-homogeneous breakdown intensity of the service station but where the actual detection of the interruption and the initiation of repair depend on the presence of a customer at the station. Hence the repair activity starts immediately if a customer is serviced at the instant of breakdown; if no customer is present the repair activity starts with a delay of

average length $1/\lambda_1$. In this model too the resume regime assumption is being made.

Elementary cycle considerations yield

$$p_o = \frac{1 + b\,\frac{\lambda_2}{\lambda_1}}{1 + \lambda_2 E(t_r) + \frac{\lambda_2}{\lambda_1}} \tag{16}$$

and

$$p_1 = \frac{\lambda_2 E(t_r) + (1 - b)\,\frac{\lambda_2}{\lambda_1}}{1 + \lambda_2 E(t_r) + \frac{\lambda_2}{\lambda_1}} \tag{17}$$

The average waiting line is derived as

$$w = \frac{\lambda_2\, E(t_r)\,(1 - \gamma_r^2)}{2\left[1 + \lambda_2 E(t_r) + \frac{\lambda_2}{\lambda_1}\right]} + \frac{\lambda_1\lambda_2 E(t_r)\;\frac{E(t_r)+(1-b)/\lambda_1}{1+\lambda_2 E(t_r)+\lambda_2/\lambda_1}\,(\gamma_r^2+1)+b^2\left[1+\lambda_2 E(t_r)\right]^2(\gamma_s^2+1)}{2\left[1 - b - b\lambda_2 E(t_r)\right]} \tag{18}$$

MODEL IV

In a few systems the station is liable to break down only while it is idle. In a sense the opposite assumption of Model II is made here. The values of p_o and p_1 are obtained as

$$p_o = \frac{1 + b\,\lambda_2\, E(t_r)}{1 + \lambda_2\, E(t_r)} \tag{19}$$

and

$$p_1 = \frac{(1-b)\lambda_2\, E(t_r)}{1+\lambda_2\, E(t_r)} \tag{20}$$

The average waiting line is given by

$$w = \frac{\frac{(1-b)\lambda_1\lambda_2 E^2(t_r)}{1+\lambda_2 E(t_r)}(\gamma_r^2+1) + b^2(\gamma_s^2+1)}{2(1-b)} \tag{21}$$

MODEL V

In this model we discard the assumption that the station breaks down from time to time in a Poissonian fashion. Rather it is assumed that a customer in service may be dissatisfied with the quality of service rendered and request that the station be adjusted to it proper level of service. Such adjustment or repair is required (at most once) by a specific customer with a probability typically depending on the number of customers served, x say, since the preceding adjustment. The random variable x is then the number of customers served between two successive repairs (with interrupted customers being counted only once). As a result of, or rather concurrent with, the adjustment the station will close down for a random time t_r. We have then

$$p_o = 1 - \frac{\lambda_1 E(t_r)}{E(x)} \tag{22}$$

and

$$p_1 = \frac{\lambda_1 E(t_r)}{E(x)} \tag{23}$$

Since the device employed in Model II - absorbing interruption time into service time - can be utilized again, it is advantageous to obtain the values of the modified busy fraction , B, and of the first two moments of the residence time, T. They are given by

$$E(T) = E(t_s) + \frac{E(t_r)}{E(x)} \tag{24}$$

$$E(T^2) = E(t_s^2) + \frac{2E(t_s)E(t_r) + E(t_r^2)}{E(x)} \tag{25}$$

and

$$B = \lambda_1 E(T) = b + p_1 \tag{26}$$

Non-saturation is expressed again by the requirement that B falls short of 1.

In analogy with Model II the expected waiting line is given by the Khintchine - Pollaczek relation with the busy fraction replaced by the modified busy fraction and the coefficient of variation relating to residence time.

$$w = \frac{B^2}{2(1-B)} (\gamma^2 + 1) = \frac{\lambda_1^2 \left[E(t_s^2) + \frac{2E(t_s)\, E(t_r) + E(t_r^2)}{E(x)} \right]}{2 \left[1 - b - \frac{\lambda_1 E(t_r)}{E(x)} \right]} \tag{27}$$

MODEL VI

On turning to queueing systems with repeat regimes we note that (at least) two different possible realizations of such regimes exist: a) On repeating service to a customer after interruption a new and independent t_s is sampled and realized (if not interrupted again); b) the nature of the customer in service determines once for all the required service time, t_s. In both cases it is, of course, assumed that t_s is sampled from a population possessing the density $f(t_s)$. In this Section we discuss case a) and Model VII is concerned with case b).

The number of interruptions, k, a customer in service undergoes is a geometrically distributed random variable; its expected value is equal to

$$E(k) = \frac{1 - L(f_s; \lambda_2)}{L(f_s; \lambda_2)} \tag{28}$$

where $L(f_s;\lambda_2)$ is the Laplace transform* of the service time density with parameter λ_2

$$L(f_s;\lambda_2) = \int_0^\infty f_s(t)\, e^{-\lambda_2 t}\, dt \tag{29}$$

*assuming, of course, that the Laplace transform of this function exists.

Let the gross service time, τ_s, be the total time a customer spends in service. This random variable is the sum of one uninterrupted and k partial (interrupted) service times. It can be shown that

$$E(\tau_s) = \frac{E(k)}{\lambda_2} \tag{30}$$

and

$$E(\tau_s^2) = \frac{2[E(k)+1]}{\lambda_2}\left(\frac{E(k)}{\lambda_2} - \frac{1}{L(f_s;\lambda_2)}\int_0^\infty t\, e^{-\lambda_2 t} f_s(t)\, dt\right) \tag{31}$$

The busy fraction of the service station is given by

$$b = \lambda_1 E(\tau_s) = \frac{\lambda_1}{\lambda_2} E(k) \tag{32}$$

and the phase probabilities are those of Model I - that is, (3) and (4) - since the identical type of breakdown - repair mechanism exists. Non-saturation is secured if the following inequality holds

$$p_o - b = \frac{1}{1+\lambda_2 E(t_r)} - \frac{\lambda_1}{\lambda_2} E(k) > 0 \tag{33}$$

The procedure outlined in the second Section leads to an expression for the expected waiting line

$$w = \frac{2b(1-b) + \lambda_1 E(t_r)\,(\gamma_r^2+1)\,p_o p_1 + b^2(\gamma_\tau+1)}{2(p_o-b)} - \frac{b}{p_o} \tag{34}$$

where γ_τ^2 is the coefficient of variation of gross service time.

If in (34) the assumption is introduced that service times are exponentially distributed an expression is derived which is identical with that of Model I with exponential service time densities. The resume regime and the present repeat regime are indistinguishable if service times are exponentially distributed.

MODEL VII

The second type of a repeat rule - the aforementioned case b) - is associated with a situation where the particular requirement for a service time is not sampled anew (and independently) whenever service is renewed to a customer after an interruption. Rather each customer has some given t_s - sampled initially from the population of service times through the agency of the density function $f_s(t)$ - and the service station has to "overcome" this value.

It can be shown that for this model the expected value of k equals

$$E(k) = \int_0^\infty e^{\lambda_2 t} f_x(t)\ dt - 1 \tag{35}$$

We note that for the integral to converge a decay stronger than $e^{-\lambda_2 t}$ is required for the density function $f_s(t)$. Gross service time, τ_s, is defined as in the previous model. Its expected value is again given by

$$E(\tau_s) = \frac{E(k)}{\lambda_2} \tag{36}$$

and the second moment can be derived as

$$E(\tau_s^2) = \frac{2}{\lambda_2^2} \int_0^\infty e^{\lambda_2 t} (e^{\lambda_2 t} - 1 - \lambda_2 t) f_s(t)\ dt \tag{37}$$

The busy fraction equals

$$b = \lambda_1 E(\tau_s) = \frac{\lambda_1}{\lambda_2} E(k) \tag{38}$$

The phase probabilities are again identical with those of Model I, but non-saturation and stationarity require a condition stronger than (33). Not only does p_o have to exceed b but the quantity

$$\int_0^\infty (e^{\lambda_2 t} - 1)^2 f_s(t)\ dt \equiv \sum_{n=1}^{\infty} \frac{\lambda_2^n (2^n - 2)}{n!} E(t_s^n) \tag{39}$$

has to be finite; it makes an appearance in the equation expressing the expected waiting line

$$w = \frac{\lambda_1 E(t_r)(\gamma_r^2 + 1)\, p_1 p_o + b^2(\gamma_\tau^2 + 1) + \dfrac{2\lambda_1^2 p_1}{\lambda_2^2 p_o} \displaystyle\int_o^\infty (e^{\lambda_2 t} - 1)^2 f_s(t)\, dt}{2(p_o - b)} \tag{40}$$

Thus, for example, detailed computation shows that for exponentially distributed service times not only the obvious inequality $b\lceil 1 + \lambda_2 E(t_r) \rceil < 1$ but also the additional inequality $2\lambda_2 E(t_s) < 1$ must be satisfied so that steady state conditions may prevail.

PREEMPTIVE PRIORITY QUEUEING

In priority queueing theory the assumption is made that it is possible to categorize the stream of incoming customers into a (usually finite) number of substreams by means of some criterion; a customer waiting or being served at the station is subject to preferences and discriminations (as compared with the simple "first come, first served" queue discipline) and these depend on the category to which he belongs. Hence it is clear that such problems fall within the general framework of the multipurpose service station theory. Indeed, it is possible to establish an immediate direct connection between the models discussed above and preemptive priority queueing where a newly arriving customer of high priority standing always displaces a low priority customer from the station. From the viewpoint of a low priority customer the arrival of a high priority customer is equivalent to a breakdown of the station. Judicious identification of priority categoryparameters and functions with those appearing in the breakdown models leads to the evaluation of quantities of interest such as average queue size of a certain customer category, etc. We shall report here on some priority queueing systems and shall endeavor to link them up with the models described above.

MODEL VIII

A station renders service to N distinct populations of customers who arrive in stationary Poisson streams with parameters $\lambda_1, \lambda_2, \ldots \lambda_n, \ldots \lambda_N$; service times will be denoted by s when discussing priority queueing and the n-th category (where n is running through the numbers $1, 2, \ldots N$) is assumed to possess a general density $f_n(s)$ with the first two moments existing and known. We use the notational convention: the higher the priority standing of a population, the lower the associated index n.

We define

$$b_n = \lambda_n \, E(s_n) \tag{41}$$

and the condition for non-saturation is

$$\sum_{n=1}^{N} b_n < 1 \tag{42}$$

On proper identification of parameters of this Model and of Model I (carried out in some detail in our 1963 paper) we obtain the expected queue length of the n-th category, a result derived previously by Miller (1960) in a completely different fashion

$$q_n = \frac{b_n}{1 - \sum_{i=1}^{n-1} b_i} + \frac{\lambda_n \sum_{i=1}^{n} \lambda_i \, E(s_i^2)}{2\left(1 - \sum_{i=1}^{n} b_i\right)\left(1 - \sum_{i=1}^{n-1} b_i\right)} \tag{43}$$

MODEL IX

If service rendered to a displaced customer is lost and must be repeated (and the new service time independently sampled) we have to utilise Model VI. After proper identification we obtain the average queue length of the n-th category as

$$q_n = \frac{b_n}{1 - \sum_{i=1}^{n-1} b_i} + \frac{\lambda_n \sum_{i=1}^{n} \left\{ \lambda_i E(\tau_i^2) \left(1 - \sum_{m=1}^{i-1} b_m\right) + \frac{2b^2}{\lambda_i} \sum_{m=1}^{i-1} b_m \right\}}{2\left(1 - \sum_{i=1}^{n} b_i\right)\left(1 - \sum_{i=1}^{n-1} b_i\right)} \tag{44}$$

where τ_i is the gross service time of a customer who belongs to i-th population. If we define

$$\Lambda_j = \sum_{i=1}^{j} \lambda_i \tag{45}$$

we obtain the average number of interruptions which an n-customer undergoes as

$$E(k_n) = \frac{1 - L(f_n; \Lambda_{n-1})}{L(f_n; \Lambda_{n-1})} \tag{46}$$

The first two moments of τ_n are given by

$$E(\tau_n) = \frac{E(k_n)}{\Lambda_{n-1}} \tag{47}$$

and

$$E(\tau_n^2) = \frac{2[E(k_n) + 1]}{\Lambda_{n-1}} \left(\frac{E(k_n)}{\Lambda_{n-1}} - \frac{\int_0^\infty s\, e^{-\Lambda_{n-1} s} f_n(s)\, ds}{L(f_n; \Lambda_{n-1})} \right) \tag{48}$$

We note that the contribution of the n-th category to the busy fraction is equal to

$$b_n = \lambda_n E(\tau_n) = \frac{\lambda_n}{\Lambda_{n-1}} E(k_n) \tag{49}$$

MODEL X

The second repeat regime deals with what was termed as case b). For a solution of the corresponding priority queueing problem we have to employ the results obtained in Model VII. The expected queue length equals

$$q_n = \frac{b_n}{1 - \sum_{i=1}^{n-1} b_i} + \frac{\lambda_n \sum_{i=1}^{n} \left\{ \lambda_i E(\tau_i^2)\left(1 - \sum_{m=1}^{i-1} b_m\right) + \frac{2\lambda_i}{\Lambda_{i-1}^2} \left(\sum_{m=1}^{i-1} b_m \right)^2 \int_0^\infty \left(e^{\Lambda_{i-1} s} - 1 \right)^2 f_i(s)\, ds \right\}}{2\left(1 - \sum_{i=1}^{n} b_i\right)\left(1 - \sum_{i=1}^{n-1} b_i\right)} \tag{50}$$

where definitions analogous to those of Model IX are used and the first two moments of gross service time are given by

$$E(\tau_n) = \frac{E(k_n)}{\lambda_{n-1}} = \frac{1}{\lambda_{n-1}} \left(\int_0^\infty e^{\lambda_{n-1} s} f_n(s)\, ds - 1 \right) \qquad (51)$$

and

$$E(\tau_n^2) = \frac{2}{\lambda_{n-1}^2} \int_0^\infty e^{\lambda_{n-1} s} (e^{\lambda_{n-1} s} - 1 - \lambda_{n-1} s)\, f_n(s)\, ds \qquad (52)$$

MODEL XI

A further priority queueing model which corresponds in some sense to Model I is the following: Two types of customers arrive at a service station. One population consists of an infinite reservoir of ordinary, low priority, potential customers who arrive in a Poisson stream with intensity ξ and are served at a rate μ. The second population consists of M extraordinary, high preemptive priority customers. Each of these extraordinary customers possesses a Poisson arrival intensity ζ. They are served by the service station at a service rate η. Both types of service times are exponentially distributed.

The quantities of interest associated with the population of high priority customers can be evaluated by considerations ordinarily made in machine interference theory (Benson and Cox (1951), Naor (1956)). The expected queue length, q, of the low priority customers can be obtained by a line of reasoning similar to that made in Model VIII.

$$q = \frac{\frac{\xi}{\mu} + \frac{p\left(M, \frac{\eta}{\zeta}\right)}{p^2\left(M, \frac{\eta}{\zeta}\right)} \frac{\xi}{\eta} \sum_{i=1}^{M} \frac{p^2\left(M-i, \frac{\eta}{\zeta}\right)}{p\left(M-i, \frac{\eta}{\zeta}\right)}}{\frac{p\left(M, \frac{\eta}{\zeta}\right)}{p\left(M, \frac{\eta}{\zeta}\right)} - \frac{\xi}{\mu}} \qquad (53)$$

where

$$p(x, \theta) \equiv e^{-\theta} \frac{\theta^x}{x!} \tag{54}$$

and

$$P(x, \theta) \equiv \sum_{y=0}^{x} p(y, \theta) \tag{55}$$

The condition for non-saturation is that the denominator of (53) is positive.

HEAD-OF-THE-LINE AND DISCRETIONARY PRIORITIES

The priority discipline assumed in Models VIII to XI is of the preemptive type. This is associated with the fact that priority queueing was considered as a particular aspect of queueing theory with a multi-purpose service station and the primary models (I - VII) which served as yardsticks were of the immediate breakdown type. Had we considered delayed breakdowns in the primary models the theory of head-of-the-line priority queueing could have been presented as an outgrowth. We shall not dwell on this here since many aspects of head-of-the-line priorities have been developed in some detail and presented (e.g. by Cobham (1954), Kesten and Runnenburg (1957), Miller (1960) and Cox and Smith (1961)). However, it appears that optimization methods pertaining to head-of-the-line disciplines may throw some light on generally optimal procedures.

Cox and Smith (1961) have derived the following important result for head-of-the-line priority queues: If each queueing customer of class n is associated with cost c_n per unit time and if customers queueing at the station are the only contributaries to the total cost function the optimal strategy of priority assignment is brought about by computing the set $\{c_n/E(s_n)\}$ and arranging the classes in descending order.

Now a head of the line priority discipline may be considered as one where review and optimal decisions occur only at certain time epochs (termination of service to a customer) whereas in preemptive priority disciplines review is con-

tinuous (in time). Hence it is suggested by Brosh and Naor (1963) that indiscriminate displacement of a low priority customer by a high priority customer is typically not an optimal policy (if review is indeed continuous). Rather the expected remaining service time of the low priority customer should play a role in the decision whether or not a displacement should occur. A discipline of this character was termed discretionary priority. In a resume regime the optimal (discretionary) rule should be the type given by Cox and Smith whereas in a repeat regime optimization rules are of a more complex character. That this is indeed the case is exemplified by the following two models (Avi-Itzhak, Brosh and Naor (1964)).

MODEL XII

Consider two populations of customers: High priority and low priority customers arrive in stationary Poisson streams with intensities λ_1 and λ_2, resp. Service times are constant and have values s_1 and s_2, respectively. Assume a high priority customer arrives while service is given to a low priority customer; if the service time already devoted to the low priority customer at the station is ϕs_2 where ϕ is an arbitrary constant ($0 \leq \phi \leq 1$) the high priority 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 high priority customer displaces the low priority customer. A resume regime is assumed.

The two expected queue lengths are

$$q_1 = b_1 + \frac{b_1^2 + b_2\lambda_1 s_2(1-\phi)^2}{2(1-b_1)} \tag{56}$$

and

$$q_2 = b_2\left(1 + \frac{\phi\, b_1}{1-b_1}\right) + \frac{b_1\lambda_2 s_1 + b_2^2}{2(1-b_1)(1-b_1-b_2)} \tag{57}$$

If we ascribe costs c_1 and c_2 (per unit time) to queueing customers of the two categories we obtain the optimal value of ϕ, ϕ^* say

$$\phi^* = 1 - \frac{c_2 s_1}{c_1 s_2} \tag{58}$$

which is meaningful if, and only if, it is in the interval (0,1). If this is not the case the priority order should be changed. The optimum value of ϕ is still given by (58) with indices interchanged. We note that (58) is precisely of the form anticipated on generalizing the result of Cox and Smith.

MODEL XIII

This Model is identical with Model XII in all but one aspect. Instead of a resume regime a repeat regime is assumed. Since service time is constant the two possible repeat regimes are identical.

Non-saturation is associated with the condition

$$\lambda_1 s_1 + \lambda_2 s_2 (1-\phi) + \frac{\lambda_2}{\lambda_1} (e^{\lambda_1 \phi s_2} - 1) < 1 \tag{59}$$

The expected queue length of high priority customers is again given by (56). However, the average number of low priority customers is represented by a rather long formula

$$q_2 = \frac{\lambda_1 \lambda_2 s_1^2 + (1-\phi)^2 \lambda_2^2 s_2^2}{2(1-\lambda_1 s_1)\left[1-\lambda_1 s_1-(1-\phi)\lambda_2 s_2 - \frac{\lambda_2}{\lambda_1}(e^{\lambda_1 \phi s_2} - 1)\right]} +$$

$$+ \frac{(1-\phi)\lambda_2 s_2\left[1-\lambda_1 s_1-(1-\phi)\lambda_2 s_2\right] + \phi \frac{\lambda_2^2}{\lambda_1} s_2 + \frac{\lambda_2}{\lambda_1}(e^{\lambda_1 \phi s_2} - 1)(1 - \frac{\lambda_2}{\lambda_1} + \lambda_2 s_2)}{1 - \lambda_1 s_1 - (1-\phi)\lambda_2 \phi_2 - \frac{\lambda_2}{\lambda_1}(e^{\lambda_1 \phi s_2} - 1)} \tag{60}$$

The selection of an optimal ϕ - assuming that costs are of the same type as in the preceding model - is not simple and no compact formula of type (58) is attainable. However, for any given set $(\lambda_1, s_1, c_1, \lambda_2, s_2, c_2)$ numerical computa-

tion of ϕ^* is not too difficult.

QUEUEING SYSTEMS WITH A REMOVABLE SERVICE STATION

There are situations where the idle fraction of a service station can be partly eliminated by closing down the service station and reopening it again at suitable times. Ordinarily a service station is not dismantled (and re-established at a later time) because it is intuitively thought that the gain obtained by dispensing with the idle fraction is not worth the loss originating from three potential sources: a) increased average queue length and queueing time, b) set-up and close-down times, and c) set-up and close-down costs.

We shall present a model based on a study by Yadin and Naor (1963) in which the effect of the following management doctrine is investigated: "Dismantle the service station, if 0 customers are in the queue; re-establish the service station, if the queue has accumulated to size R".

MODEL XIV

We assume customers to arrive in a Poisson stream with intensity λ_1; the station is closed down and set up from time to time and these processes consume time; set-up times, t_α, and close-down times t_β are statistically distributed, and so are, of course, service times, t_s. We assume that these random variables possess (arbitrary) densities; the existence and knowledge of the first two moments only will be needed for the service and set-up times whereas the Laplace transform of the close-down time density must be specifiable. Further, it is assumed that the physical elimination of the station takes place only with the termination of the closing-down process; that on arrival of the new customer the closing-down process is immediately interrupted and the server is instantaneously placed at his disposition; and that the time already spent on closing-down in the present attempt is of no avail for the next closing down occasion.

We can distinguish four phases in the cycle and the associated probabilities are the following

$$\Pr\{\text{station inactive}\} = \frac{R(1-b)}{R+\alpha+\beta} \tag{61}$$

$$\Pr\{\text{set-up period}\} = \frac{\alpha(1-b)}{R+\alpha+\beta} \tag{62}$$

$$\Pr\{\text{station active}\} = b \tag{63}$$

$$\Pr\{\text{close-down period}\} = \frac{\beta(1-b)}{R+\alpha+\beta} \tag{64}$$

where α and β are defined as

$$\alpha = \lambda_1 E(t_\alpha) \tag{65}$$

and

$$\beta = \frac{1-L(f_\beta, \lambda_1)}{L(f_\beta, \lambda_1)} \tag{66}$$

The physical interpretation of α and β is simple: α is the expected number of customers who arrive at the station during a set-up period; β is the expected number of customers who arrive at the station during a (gross) close-down period.

The expected queue length as a function of the basic, natural parameters $(b, \alpha, \beta, \gamma_s, \gamma_\alpha)$ and the controllable variable R is represented as

$$q = b + \frac{b^2}{1-b} \cdot \frac{1+\gamma_s^2}{2} + \frac{R}{R+\alpha+\beta} \cdot \frac{R-1}{2} + \frac{\alpha}{R+\alpha+\beta}\left[R + \alpha \frac{1+\gamma_\alpha^2}{2}\right] \tag{67}$$

The first two terms (Khintchine-Pollaczek relation) describe the queue length under conditions of permanent service station availability. The sum of the third and fourth terms is the contribution to average queue length as a result of the intermittent operation of the service station.

Reasonable cost assumptions include a) a cost C_c due to each queueing customer (per unit time), b) cycle expenditure C_T, that is the sum of the set-up and close-down costs, and c) savings C_s (per unit time) of the inactive station.

For the application of the management doctrine (0, R) to be advantageous, at least one integral value of $R(\geq 1)$ must exist such that the value of the total cost function is lower than for a permanently available station. It can be shown that the criterion for this is

$$C_s \left[R + (1 - \psi_\alpha)A + (1 - \psi_\beta)\beta \right] > \frac{C_c}{1-b} \left[\frac{R(R-1)}{2} + \alpha R + \alpha \frac{1 + \gamma_\alpha^2}{2} \right] + \lambda C_T \tag{68}$$

where ψ_α and ψ_β are the fractions of set-up and close-down time, during which the ordinary station expenditure is continued. If (at least) one value of R exists which satisfies (68) the optimal R, R* say, is given (approximately) by

$$R^* = \left\{ 2(1-b) \left[(\psi_\alpha \alpha + \psi_\beta \beta) \frac{C_s}{C_c} + \frac{\lambda C_T}{C_c} \right] + \alpha(1 + \alpha \gamma_\alpha^2) + \beta(1 + \beta) \right\}^{\frac{1}{2}} - (\alpha + \beta) \tag{69}$$

This is closely related to square root formulas appearing in inventory theory.

CONCLUSION

The models reviewed in this paper represent a small (but possibly useful) sample of a multitude of situations. The approach used is to consider station breakdowns, interruptions, priority queueing etc. as different manifestations of a multi-purpose station performing various missions.

REFERENCES

Avi-Itzhak, B., and Naor, P. "On a Problem of Preemptive Priority Queueing," Opns. Res. 9, 664-672 (1961).

Avi-Itzhak, B., "Preemptive Repeat Priority Queues as a Special Case of the Multi-Purpose Server Problem, I and II," to be published in Opns. Res. 11 , (1963).

Avi-Itzhak, B. and Naor, P., "Some Queueing Problems with the Service Station Subject to Breakdown," to be published in Opns. Res. 11 , (1963).

Avi-Itzhak, B. "A Queueing Model with a Deteriorating Service Station," (1963) to be published; available in mimeographed form from the Dept. of Industrial and Management Engineering, Technion, Israel Institute of Technology, Haifa, Israel.

Avi-Itzhak, B., Brosh, I., and Naor, P., "On Discretionary Priority Queueing," to be published in Zeitschrift für Angewandte Mathematik und Mechanik, 44 , (1964). Available in mimeographed form from the Department of Statistics, University of North Carolina, Chapel Hill, N. C.

Benson, F. and Cox, D. R., "The Productivity of Machines Requiring Attention at Random Intervals," J. R. Statist. Soc. B, 14 , 200-210 (1951).

Brosh, I., and Naor, P., "On Optimal Disciplines in Priority Queueing," paper to be read at the 34th Session of the International Statistical Institute, Ottawa, Ontario, Canada, August 1963. Available in mimeographed form from the Department of Statistics, University of North Carolina, Chapel Hill, N. C.

Cobham, A., "Priority Assignment in Waiting Line Problems," Opns, Res. 2 , 70-76 (1954).

Cox, D. R. and Smith, W. L., "On the Superposition of Renewal Processes," Biometrika, 41, 91-99 (1954).

Cox, D. R. and Smith, W. L., "Queues", Methuen and Co., Ltd. London, John Wiley and Sons, Inc., New York (1961).

Gaver, D. P., Jr., "A Waiting Line with Interrupted Service, Including Priorities," J. R. Statist. Soc. B, 24 , 73-90 (1962).

Heathcote, C. R., "Preemptive Priority Queueing," Biometrika, 48, 57-63 (1961).

Keilson, Julian, "Queues Subject to Service Interruption," Ann. Math. Stat. 33, 1314-1322 (1962).

Kesten, H. and Runnenburg, J. Th., "Priority in Waiting Line Problems, I and II," Proc. Akad. Wet. Amst. A, 60, 312-324, 325-336 (1957).

Little, John D. C., "A Proof for the Queueing Formula: $L = \lambda W$" Opns. Res. 9, 383-387, (1961).

Miller, R. G., "Priority Queues," Ann. Math. Stat. 31, 86-103, (1960).

Naor, P., "On Machine Interference," J. R. Statist. Soc. B, 18, 280-287, (1956).

Naor, P., "Some Problems of Machine Interference," Proc. First Inter. Conf. on Operational Res. 147-164, English Universities Press, London (1957).

Palm, C., "Intensitätsschwankungen im Fernsprechverkehr," Ericsson Technics, 44, 1 - 189 (1943).

v. Smoluchowski, M., "Molekulartheoretische Studien über Umkehr thermodynamisch irreversibler Vorgänge und über Wiederkehr abnormaler Zustände," S. B. Akad. Wiss. Wien (IIa) 124, 339-368 (1915).

Thiruvengadam, K., "Queueing with Breakdowns," Opns. Res. 11, 62-71, (1963).

White, Harrison and Christie, Lee S., "Queueing with Preemptive Priorities or with Breakdown," Opns. Res. 6, 79-95 (1958).

Yadin, M., and Naor, P., "Queueing Systems with a Removable Service Station," to be published, available in mimeographed form from the Department of Statistics, University of North Carolina, Chapel Hill, N. C. (1963).