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

QUEUEING SYSTEMS WITH A REMOVABLE SERVICE STATION

by

M. Yadin and P. Naor
Technion, Israel Institute of Technology, Haifa, Israel
and
University of North Carolina, Chapel Hill, N. C.

February 1963

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. 353

QUEUEING SYSTEMS WITH A REMOVABLE SERVICE STATION

by

M. Yadin and P. Naor

Technion, Israel Institute of Technology, Haifa, Israel

and

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

ABSTRACT

A single server queueing system with constant Poisson input is considered and the partial elimination of the station's idle fraction is envisaged by intermittent close-down and set-up. The rule pertaining to the dismantling and re-establishing of the service station - the management doctrine - is based on the instantaneous size of the queue, but these processes are assumed to consume time. Operating characteristics of such systems - in particular, average queue length and queueing time - are evaluated. A cost structure is superimposed on the system and optimisation procedures are outlined. The close relationship with a) priority queueing and b) storage models is pointed out.

QUEUEING SYSTEMS WITH A REMOVABLE SERVICE STATION[≁]

by

M. Yadin and P. Naor

Technion, Israel Institute of Technology
Haifa, Israel

and

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

= =

I. Introduction

It is a well-known result that the attainment of steady state in a queueing system depends on the existence of "excess service capacity". In a simple single-server situation with a (statistically) steady input of customers and a (statistically) steady service rate this condition may be expressed by the statement: "The average time elapsing between the arrivals of two successive customers must exceed the average service time".[*] Indeed it is intuitively obvious how to generalize this statement in practical situations where the prior assumptions of steady input and service rate (and even of permanent availability of the station) are not

[*] Two explanatory and qualifying remarks are not out of place here: a) "Arrival" of a customer as envisaged in the above statement entails his joining the queue and staying in it until completion of service. Customers who do not join or who leave prematurely must be discounted in the calculation of the average interarrival time. b) If both arrivals and service times follow a completely deterministic (and periodic) pattern (which, after a fashion, is a special degenerate statistical situation) the "average" interarrival time may even be equal to the "average" service time. Obviously no steady state in a statistical sense is attained; rather queue length is a periodic function of time.

[≁] 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.

realised and furthermore it is not difficult to derive the appropriate extension for the many server case. The basic outcome always obtained is the following: For steady state conditions to prevail the situation in the queueing system must be such that during a fraction of time the server(s) is (are) not engaged in rendering service.

Furthermore it appears that in many practical situations this idle fraction of the server's time must be quite appreciable if the ordinarily important indices of performance (such as: average queue length, expected waiting time, etc.) are to be kept within reasonable bounds. Thus, for instance, in the single server case with (stationary) Poisson arrivals and generally distributed service times the expected waiting line (exclusive of the customer in service) is given by the expression $C(1-K)^2/K$ where K is the idle fraction and C a constant (usually of the order $0.5 \leq C \leq 1$) depending on the coefficient of variation of the service time distribution. If it is desired that approximately one customer (on the average) should be waiting for the service the required idle fraction of the server must be of the order 0.30 - 0.40, certainly no small number.

If the establishment of service stations is considered a fixed expense (that is one which is occurring continuously in time at a constant rate) the problem of utilising the server's idle fraction arises. Attention was paid to this problem as early as 1952 in a study by Benson[1] who presented an analysis of the ancillary duties of an operator (that is: duties to be carried out during his idle fraction) compatible with his main function within a machine interference context. A further development of major importance in this connection was the explicit formulation of priority queueing problems by Cobham[2]. From the viewpoint[*] of the high priority customer service rendered to a low priority customer is tantamount to

[*] It is interesting to note that the alternative viewpoint - that of the low priority customer - leads to models of queueing systems with the service station subject to random breakdowns.

utilisation of the idle fraction.

Now there is an alternative approach to the problem of the idle fraction: its elimination by closing down the service station and reopening it again at suitable times. For systems with variable customer inputs - in particular, when average input is a periodic function of time - this device must have been used since times immemorial. However, it is not immediately obvious that this doctrine may be applicable,useful and (sometimes) optimal even with a statistically steady input of customers. Clearly we are not concerned here with a situation where service stations may be established and dismantled without financial (or other) penalty or where servers may be hired and fired without impunity; in that unrealistic context the above doctrine is trivially optimal.

Ordinarily a service station is not dismantled (and re-established at a later time) precisely 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. The purpose of this communication is to present an analysis of the single server case (or rather single-zero server case) under assumptions of quite general character. To be more precise, the effect of an (0, R) doctrine ("dismantle the service station, if 0 customers are in the queue; re-establish the service station, if the queue has accumulated to size R") on a queueing system is investigated.

A reasonable degree of generality in our approach is attained by incorporating statistically distributed set-up and close-down times in the model under study. Furthermore, very few assumptions will be made with respect to the service time, set-up time and close-down time distributions. They will be assumed to possess arbitrary functions (each having a density); 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. Two specialised features form part of our model: First, customers are assumed to arrive in a stationary Poisson stream; this assumption is common and not unrealistic in the sense that many empirical "real life" situations are observed to behave in a fashion compatible with it. Next, a statement must be made regarding the contingency of a customer arriving at the station while it is being closed down. Several alternatives with respect to both external constraints and possible actions* of the server are open. The one chosen in this study is that the physical elimination of the server 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

*In terms of priority queueing the closing-down of the station may be interpreted as (a part of) a service time rendered to a low priority customer. The new arrival of a customer (in the terminology of this paper) is equivalent to the arrival of a high priority customer (in priority queueing terminology). Several alternative rules of action present themselves here: a) preemptive priority rule, b) head of the line priority rule, and c) discretionary priority[3] rule (incorporating both rules a) and b) as special, extreme cases) which would regard the possible interruption or non-interruption of the closing-down process (as a function of closing-down time already elapsed) to be an extension of the management doctrine. Our actual choice in this study is alternative a) and within this context a "repeat" situation is assumed to prevail. This particular area within priority queueing has been recently explored by Avi-Itzhak[4] and one important result of his was utilised in this study.

time already spent on closing-down in the present attempt is of no avail for the next closing-down occasion .

In effect the above outline represents a complete, qualitative description of the model under study. We take note that the cost structure does not form an integral part of the model. Rather, the model is composed of (assumptions regarding) processes occurring in time, natural constraints, and imposable (or adjustable) constraints - the latter being the so-called management doctrine. Now, of course, this is quite characteristic of many queueing, inventory and other storage problems whereas in the typical programming problems analysis in cost terms plays a vital part from the very beginning of the solution procedure. In this study, then, what is termed a solution is a set of more or less compact relations expressing the operating characteristics of interest in terms of the "natural" parameters and of the management doctrine.

The cost structure is then superimposed on the solution and optimisation is carried out. In view of the fact that for any given queueing (or similar) problem the cost structure is not uniquely defined we have not pursued optimisation too far. Thus the particular cost structure chosen in this paper and the ensuing analysis should be considered as a simple sample of possible approaches to optimisation and not as a vital stage in the development of the study.

II. Some Definitions and General Relations

We recapitulate and formalize the model under study: Customers arrive at a single service station in a stationary Poisson stream with parameter λ; the service station (when active) renders and completes service to a customer during a random time interval t_s. The service times devoted to different customers are mutually independent and identically distributed; their common distribution function is $F_s(t)$ and all we assume is that the first two moments - $E_s(t)$ and

$E_s(t^2)$ - exist and are known. If the service station is incapacitated and a decision is taken to reactivate it, a random time interval - the set-up time, t_α - will elapse before the service activity is actually renewed. The distribution function pertaining to this set-up time is $F_\alpha(t)$ and again our sole requirements are the existence and knowledge of the first ~~two~~ moments, $E_\alpha(t)$ and $E_\alpha(t^2)$. A decision to close-down the service station is reached from time to time - in the present study this occurs if, and only if, no customers are at the station. This decision, too, cannot be put into practice instantaneously and close-down time, t_β, is a random variable with distribution function $F_\beta(t)$ and density function $f_\beta(t)$. For reasons which will become obvious at a slightly later stage we shall assume that the Laplace transform (with parameter λ) of the density function exists. This is equivalent to assuming that all moments of the distribution exist and, of course, is less general than the assumptions regarding $F_s(t)$ and $F_\alpha(t)$.

Without formal proof we equate non-saturation with the existence of steady state conditions. Thus the only type of situations investigated here is associated with the inequality

$$E(t_s) < \frac{1}{\lambda} \qquad (1)$$

which enables us to define the concept of a busy fraction in an unambiguous fashion

$$b = \lambda\, E(t_s) < 1 \qquad (2)$$

It is convenient to use the concept of busy period in our study. This is the period that elapses from the time at which the previously disengaged service station starts to render service to a customer until the moment at which the last customer (out of a whole train of customers who arrived while service was given to earlier customers) is cleared from the station. Thus in the simple queueing situations the service station experiences cycles consisting of two phases: a) the busy period and b) the clear period. Let the average duration of these periods be denoted by T_b and T_c, respectively. By virtue of the properties of the Poisson arrival process the expected length of the clear period equals

$$T_c = \frac{1}{\lambda} \tag{3}$$

Furthermore by definition of the concept of the busy fraction the following must hold

$$\frac{T_b}{T_b + T_c} = b \tag{4}$$

On combining (2), (3) and (4) we obtain

$$T_b = \frac{b}{\lambda(1-b)} = \frac{E(t_s)}{1-b} \tag{5}$$

This relation has the following physical interpretation. The station completes service to a customer within an average period $E(t_s)$. However, servicing a customer is associated with aftereffects: while the service process is going on (and the station is engaged) new customers may arrive and, in turn, while these "second generation" customers are being dealt with, "third generation" customers may arrive etc. Service to a whole train of customers as described above is completed only after an average period $E(t_s)/1-b$ which exceeds, of course, the average service time devoted to one customer. It is important to note that this formula depends on the assumption of Poisson arrivals since use is made of (3) in its derivation.

We are now in a position to analyze the four-phase cycle of the model under-study. It is convenient to start at the moment at which the closing-down of the station has been successfully concluded. We have no customers at the station and the station is incapable of rendering service to any customers who may arrive during the present phase. We recollect that no attempt is made to reactivate the station until R customers have assembled. Obviously, the average assemblying time is equal to R/λ.

The second phase is the set-up time $E_\alpha(t)$. Clearly a number of customers arrives at the station during the set-up time and joins the queue of R customers already present there. Let the average number of such additional arrivals be denoted by α; this quantity may be evaluated as

$$\alpha = \int_o^\infty \sum_{i=o}^\infty i\, e^{-\lambda t} \frac{(\lambda t)^i}{i!} d\, F_\alpha(t) = \int_o^\infty \lambda t\, d\, F_\alpha(t) = \lambda\, E_\alpha(t) \qquad (6)$$

At the beginning of the third phase there are (on the average) $R + \alpha$ customers present at the station which is now ready to render service. From relation (5) we know that $E(t_s)/1\text{-}b$ time units are required to complete service to <u>one</u> customer and to his potential second, third, etc. generations. Since $R + \alpha$ customers request service at the beginning of the third phase it will last (on the average) for a time $(R + \alpha)\, E(t_s)/1\text{-}b$.

The evaluation of the average length of the fourth phase is more complex. At the beginning of this phase no customer is present, the station is capable of rendering service if required, and the close-down process is about to start. We have then <u>two</u> concurrent and competing random processes going on which must terminate either in completion of close-down or in the arrival of a new customer who seizes the service station and - as explained in the Introduction - cancels all close-down work carried out hitherto. Let us denote the probability of terminating the fourth phase at this stage (by completion of close-down) by π. This quantity is easily seen to be the Laplace transform of $f_\beta(t)$ with respect to the parameter λ.

$$\pi = \int_0^\infty e^{-\lambda t} \; f_\beta(t) \, dt = \mathcal{L}(f_\beta, \lambda) \qquad (7)$$

The probability of the complementary event (arrival of a new customer before completion of close-down) occurring is equal, of course, to $1 - \mathcal{L}(f_\beta,\lambda)$.

The number of unsuccessful attempts (within one cycle) to complete close-down is a geometrically distributed random variable; its frequency function is given by

$$p_n = \pi(1-\pi)^n \qquad (n = 0,1,2,\dots) \qquad (8)$$

Now this random variable is also equal to the number of customer's arrivals (during one cycle) while the close-down process is taking place. Hence the expected value of n is a quantity analogous to α; it will be denoted by β.

$$\beta = E(n) = \frac{1-\pi}{\pi} = \frac{1 - \mathcal{L}(f_\beta,\lambda)}{\mathcal{L}(f_\beta,\lambda)} \qquad (9)$$

Within the fourth phase of a cycle we have (on the average) β unsuccessful attempts to complete close-down and one successful attempt. For our analysis it is necessary to establish the average gross close-down time per cycle, that is, the sum total of a) the β partial, interrupted close-down times and b) the final successful close-down time. Avi-Itzhak[4] has solved this problem in the context of preemptive repeat priority queueing and the line of reasoning we shall employ follows one of his original proofs.

We noted before that two concurrent and competing random processes take place while the station is being closed down; one is the close-down process itself and the other is the customer arrival process. We wish to obtain the conditional density functions $f_1(t)$ and $f_2(t)$ of the duration of this competition given 1) that it ended in the arrival of a customer, and 2) that it ended in completion of close-down. By elementary and fundamental considerations of conditional probabilities we arrive at

$$f_1(t) = \frac{1}{1-\pi} \ \lambda \ e^{-\lambda t} \left[1 - F_\beta(t) \right] \tag{10}$$

and

$$f_2(t) = \frac{1}{\pi} \ f_\beta(t) \ e^{-\lambda t} \tag{11}$$

The expected gross close-down time, $E_\beta(t_g)$, equals

$$E_\beta(t_g) = \beta \ E_1(t) + E_2(t) =$$

$$= \frac{1-\pi}{\pi} \int_0^\infty \frac{\lambda t}{1-\pi} \ e^{-\lambda t} \left[1 - F_\beta(t) \right] dt + \int_0^\infty \frac{t}{\pi} e^{-\lambda t} f_\beta(t) \ dt =$$

$$= \frac{1}{\pi} \left[\int_0^\infty \int_t^\infty \lambda t e^{-\lambda t} \ f_\beta(T) dT dt + \int_0^\infty t e^{-\lambda t} \ f_\beta(t) \ dt \right] =$$

$$= \frac{1}{\pi} \left[\int_0^\infty \int_0^t \lambda T e^{-\lambda T} \ dT \ f_\beta(t) \ dt + \int_0^\infty t e^{-\lambda t} \ f_\beta(t) \ dt \right] =$$

$$= \frac{1}{\lambda \ \pi} \left[\int_0^\infty (1 - e^{-\lambda t} - \lambda t e^{-\lambda t}) \ f_\beta(t) dt + \int_0 \lambda t e^{-\lambda t} \ f_\beta(t) \ dt \right] =$$

$$= \frac{1}{\lambda \pi} \int_0^\infty (1 - e^{-\lambda t}) f_\beta(t) dt = \frac{1}{\lambda \pi} \left[1 - \mathcal{L}(f_\beta, \lambda) \right] = \frac{1}{\lambda} \cdot \frac{1-\pi}{\pi} =$$

$$= \frac{\beta}{\lambda} \tag{12}$$

Avi-Itzhak's formula has an intuitive interpretation: the average number of customers who arrive during the gross close-down time is equal to the product of the arrival intensity (number of customers arriving in unit time) and the average duration of gross close-down time. However, in this case the intuitive interpretation is of the hindsight variety. It is by no means obvious that the particular combination of components making up the gross close-down time - some of them originating in the f_1-density and one taken from the f_2-density - should yield the final simple relation (12).

The fourth phase consists of the gross close-down time and of β busy periods each starting with one customer at the station. Hence the average duration of the fourth phase is equal to $\beta/\lambda + \beta\, E(t_s)/1\text{-}b$.

A relation of cardinal importance - expressing the average duration of a cycle, T, as a function of the basic model parameters - is obtained on summing the average lengths of the four phases

$$T = \frac{R}{\lambda} + \frac{\alpha}{\lambda} + \frac{(R+\alpha)E(t_s)}{1-b} + \left(\frac{\beta}{\lambda} + \frac{\beta\ E(t_s)}{1-b} \right) =$$

$$= (R + \alpha + \beta) \left(\frac{1}{\lambda} + \frac{E(t_s)}{1-b} \right) = \frac{R + \alpha + \beta}{\lambda(1-b)} \qquad (13)$$

The probabilities of the system being in the various phases and subphases are obtained as

$$\Pr \left\{ \text{system in first phase} \right\} = \frac{R}{\lambda T} = \frac{R(1-b)}{R + \alpha + \beta} \qquad (14)$$

$$\Pr \left\{ \text{system in second phase} \right\} = \frac{\alpha}{\lambda T} = \frac{\alpha(1-b)}{R + \alpha + \beta} \qquad (15)$$

$$\Pr \left\{ \text{system in third phase} \right\} = \frac{(R+\alpha)\ E(t_s)}{(1-b)T} =$$

$$= \frac{(R+\alpha)b}{R + \alpha + \beta} \qquad (16)$$

$$\Pr \left\{ \text{system in fourth phase, close-down subphase} \right\} =$$

$$= \frac{\beta}{\lambda T} = \frac{\beta(1-b)}{R + \alpha + \beta} \qquad (17)$$

$$\Pr \left\{ \text{system in fourth phase, active station subphase} \right\} =$$

$$= \frac{\beta\ E(t_s)}{(1-b)T} = \frac{\beta\ b}{R + \alpha + \beta} \qquad (18)$$

Combinations of these probabilities yield further results of interest. Thus, for instance, adding up (17) and (18) generates the (unconditional) probability of the system being in the fourth phase - $\beta/R+\alpha+\beta$. On addition of (14), (15) and (17) the classical* idle fraction (1-b) is obtained. Further examples are easily furnished.

III. Average Queue Length and Queueing Time

To evaluate the average queue length and queueing time further refinements in notation have to be introduced. Let $E_{ki}(k = 0,1; i = 0,1,2,...)$ describe the state of the system in which k available service stations and i queueing customers (including the one in service) are present; p_{ki} is the probability associated with state E_{ki}. Connection between the cycle approach (of the preceding Section) to the analysis of the model and the present approach is established by noting that $\bigcup_{i=0}^{R-1} E_{Oi}$ is that collection of states through which the system passes in the first phase. Since this "passing through" comes about by Poisson jumps in the direction of increasing i all states of this union are equiprobable. Hence by (14)

$$p_{oi} = \frac{1-b}{R + \alpha + \beta} \qquad (i = 0,1,... R-1) \qquad (19)$$

Furthermore the union of states $\bigcup_{i=R}^{\infty} E_{oi}$ corresponds to the second phase of the cycle. Thus (15) is transformed into

$$\sum_{i=R}^{\infty} p_{oi} = \frac{\alpha(1-b)}{R + \alpha + \beta} \qquad (20)$$

$\bigcup_{i=1}^{\infty} E_{li}$ is that collection of states during which the station is busy. The asso-

*What is precisely meant by "idle fraction" in the present model depends on additional assumptions. This will be discussed when cost factors are introduced.

ciated probability is readily seen to be

$$\sum_{i=1}^{\infty} p_{1i} = b \qquad (21)$$

This result may also be obtained by adding (16) and (18) .

It is important to recollect the following: Assume that a customer (originating, of course, from a Poisson stream) arrives at the service station. Let this service station be engaged in some activity - e.g., serving a preceding customer, or being set-up - with which the newly arrived customer is not going to interfere. If these conditions prevail the time interval between the arrival of the customer and the termination of the activity is the "forward delay" (Cox and Smith[5]) of the activity duration. The expected value of this forward delay, τ , bears a simple relationship to the first two moments of the original random variable t (activity duration), to wit

$$E(\tau) = \frac{E(t^2)}{2\,E(t)} = E(t)\;\frac{1+\gamma^2}{2} \qquad (22)$$

where γ is the coefficient of variation of t. A customer who arrives during $\bigcup_{i=R}^{\infty} E_{oi}$ or $\bigcup_{i=1}^{\infty} E_{1i}$ encounters the service station "in the midst" of an activity. Hence the average queueing time of such a customer will contain terms $E(\tau_\alpha)$ or $E(\tau_s)$, respectively.

We proceed now to the detailed analysis of average queueing times for customers arriving at different epochs.

First we consider a customer who arrives during a state $E_{oi}(0 \leq i \leq R-1)$. His average queueing time is made up of three parts - $(R-1-i)/\lambda+E(t_\alpha)+(i+1)E(t_s)$. The first term is representative of the remaining assemblying time, the second term is the set-up time of the station, and the third term is the service time devoted to the i preceding customers and to himself.

Let a customer enter the system during a state $E_{oi}(R \leq i < \infty)$. By virtue of his random arrival "in the midst" of setting - up the station one contribution that is made to his average queueing time is $E(\tau_\alpha)$* . A second contribution is identical with the last term in the previous case - $(i+1)E(t_s)$.

The contingency of a customer showing up during $E_{1i}(1 \leq i < \infty)$ is analyzed next. Here again queueing time is made up of two components. First, service must be completed to the first customer (the one in service). This average completion time equals $E(\tau_s)$.** Secondly, the service station must take care of $i-1$ preceding customers and the new customer himself. The average time contribution of this component equals $i\, E(t_s)$.

Finally the customer may make his appearance during state E_{10}. This immediately interrupts the close-down process, the customer is promptly served and the average queueing time equals $E(t_s)$ in this case.

We are now in a position to establish the average queueing time, Θ_q, as

$$\Theta_q = \sum_{i=0}^{R-1} p_{oi} \left[(R-1-i) \frac{1}{\lambda} + E(t_\alpha) + (i+1)\, E(t_s) \right] +$$

$$+ \sum_{i=R}^{\infty} p_{oi} \left[E(\tau_\alpha) + (i+1)\, E(t_s) \right] +$$

$$+ \sum_{i=1}^{\infty} p_{1i} \left[E(\tau_s) + i\, E(t_s) \right] + p_{10}\, E(t_s) \qquad (23)$$

*If it is not the average contribution (over all appropriate i) in which we are interested extreme care must be taken in the analysis, since typically $E(\tau_\alpha | i)$ is not equal to $E(\tau_\alpha)$. The shortcut we use in this paper (suppressing terms of the type $E(\tau_\alpha | i)$) is based on the identity $E(\tau_\alpha) \sum_{i=R}^{\infty} p_{oi} = \sum_{i=R}^{\infty} p_{oi}\, E(\tau_\alpha | i)$.

**See previous footnote. The argument leading to the direct use of $E(\tau_s)$ instead of starting out with $E(\tau_s | i)$ is completely analogous to that relating to $E(\tau_\alpha)$.

Insertion of (19), (20), (21), and (22) transforms (23) into

$$\Theta_q = \frac{1}{\lambda} \left[(q+1-b)b+b^2 \; \frac{1+\gamma_s^2}{2} + \frac{R(R-1)(1-b)}{2(R+\alpha+\beta)} + \right.$$

$$\left. + \frac{\alpha R(1-b)}{R+\alpha+\beta} + \frac{\alpha^2(1-b)}{R+\alpha+\beta} \cdot \frac{1+\gamma_\alpha^2}{2} \right] \qquad (24)$$

Now there exists a fundamental connection* between q and Θ_q

$$q = \lambda \, \Theta_q \qquad (25)$$

Combination of (24) and (25) yields an expression for q as a function of the basic, natural parameters (b, α, β, γ_s, γ_α) and the controllable variable R.

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

$$(26)$$

The average queue length as represented by the right hand side of (26) is made up of several terms. The first two terms (which are equivalent to the Khintchine - Pollaczek relation) describe the queue length under conditions of <u>permanent</u> service station availability. The sum of the third and fourth terms is the contribution to average queue length as a result of the <u>intermittent</u> operation of the service station. Roughly speaking, the magnitude of the third term depends primarily on the chosen value of R whereas the fourth term is more representative of the impact of the set-up parameters; indeed, if there is no set-up time the fourth term disappears.

*Little[6] made a rigorous analysis and stated conditions under which relations of type (25) hold. He showed, in effect, that these relations may be established if, and only if, the average arrival intensity is equal to the average departure intensity. Clearly, this is the case in the present study.

IV. Cost Structure and Optimization

The cost function which we construct and consider here is of the simplest possible type: The contributions of the various components to the total cost are assumed to be linear with respect to their average values. Essentially there exist three such components contributing to the additional (positive or negative) cost, ΔC_R, caused by the application of the doctrine $(0,R)$.

The first of these originates from increased queue length. For convenience we define the sum of the last two terms in (26) - the additional queue length due to the intermittent operation of the service station - as Δq_R.

$$\Delta q_R = \frac{1}{R + \alpha + \beta} \left[\frac{R(R-1)}{2} + \alpha\left(R + \alpha \frac{1 + \gamma_\alpha^2}{2}\right) \right] \qquad (27)$$

If the presence of a single waiting customer at the station costs C_c per unit time this term is associated with a contribution to the total cost function amounting to $C_c \Delta q_R$.

Secondly we are concerned with cycle expenditure, i.e., the sum of the set-up and close-down costs. If this cost (per cycle) is designated as C_T the contribution per unit time equals C_T/T.

Finally we must consider the negative cost - i.e., savings - brought about by the partial elimination of the service station during its idle fraction. Let the fractions of set-up and close-down time*, during which the ordinary station expenditure is continued, be denoted as ψ_α and ψ_β, respectively. If $(1 - \phi)$ is

*It is, of course, possible to exclude costs pertaining to these times from station expenditure, incorporate them in cycle expenditure, and reach an alternative (but equivalent formulation of the total cost function. We have chosen the above representation to gain further insight into the cost structure.

that fraction of time during which the station incurs expenditure we have

$$(1 - \phi) = b + (1-b)\ \frac{\psi_\alpha \alpha + \psi_\beta \beta}{R + \alpha + \beta} \tag{28}$$

and

$$\phi = \frac{1-b}{R+\alpha+\beta}\ \left[R + (1 - \psi_\alpha)\alpha + (1 - \psi_\beta)\beta \right] \tag{29}$$

The contribution of this term equals $- C_s \phi$ where C_s is the cost (per unit time) of maintaining a service station.

The total (additional) cost function, ΔC_R, is obtained on summing these three cost contributions

$$\Delta C_R = C_c \Delta q_R + C_T/T - C_s \phi =$$

$$= \frac{1}{R + \alpha + \beta} \left\{ C_c \left[\frac{R(R-1)}{2} + \alpha(R + \alpha \frac{1 + \gamma_\alpha^2}{2}) \right] + \right.$$

$$\left. + C_T \lambda(1-b) - C_s(1-b)\left[R + (1 - \psi_\alpha)\alpha + (1 - \psi_\beta)\beta \right] \right\} \tag{30}$$

Clearly for the application of the doctrine (0,R) to be advantageous, at least one integral value of $R(\geq 1)$ must exist such that ΔC_R is a negative quantity. Hence relation (30) leads to the criterion

$$C_s\left[R + (1 - \psi_\alpha)\alpha + (1 - \psi_\beta)\beta \right] > \frac{C_c}{1-b} \left[\frac{R(R-1)}{2} + \alpha R + \alpha \frac{1 + \gamma_\alpha^2}{2} \right] + \lambda C_T \tag{31}$$

In practice the following procedure is suggested for the attainment of the optimum solution;

Equation (30) is differentiated with respect to R and the derivative is set equal to zero. This leads to the quadractic equation

$$R^2+2(\alpha+\beta)R-\left\{(\gamma_\alpha^2-1)\alpha^2+(1-2\beta)\alpha+\beta+2(1-b)\left[(\psi_\alpha\alpha+\psi_\beta\beta)\frac{C_s}{C_c}+\lambda\frac{C_T}{C_c}\right]\right\}=0 \qquad (32)$$

One of its solutions, R^*, is (approximately) feasible, to wit

$$R^*=\left\{2(1-b)\left[(\psi_\alpha\alpha+\psi_\beta\beta)\frac{C_s}{C_c}+\lambda\frac{C_T}{C_c}\right]+\alpha(1+\alpha\gamma_\alpha^2)+\beta(1+\beta)\right\}^{\frac{1}{2}}-(\alpha+\beta) \qquad (33)$$

R^* is rounded down and up to the nearest positive integers $[R^*]$ and $[R^*]+1$; the (provisionally) optimal solution is obtained by comparing $\Delta C_{[R^*]}$ and $\Delta C_{[R^*+1]}$. Finally this solution is considered against the alternative of permanent station availability by means of criterion (31).

It is interesting to note that in the absence of set-up and close-down times relation (33) is transformed into a square root formula very similar to those making their appearance in elementary inventory theory.

$$R^*=\sqrt{2(1-b)\lambda\frac{C_T}{C_c}} \qquad (34)$$

V. Conclusion

There are numerous situations - industrial and otherwise - which are well described by the model presented in this study.

Consider, for instance, a large set of machines each of which is liable to break down from time to time in some random fashion and to require the attention of a maintenance crew. It may be quite expensive to call in the maintenance crew as soon as a piece of equipment has failed and frequently the policy of having periodic (or intermittent) inspection by the crew, independent of the number of inactive machines in the system may be far from optimal. The procedures outlined in this study may well be applicable to such a situation.

Another example from a different area of activity is the following: An office collects information and is responsible for its transmission to various addresses by cable*, say. The information arrives at the office in a Poisson stream but it may not be economical to permanently hold the transmitting equipment in an expensive stand-by state. It may be possible to attain an optimal mode of operation by an analysis closely related to that of the present paper.

Vehicle-actuated traffic lights belong to the class of situations described by our model. A part of their theory is completely covered by the results obtained in this investigation.

For a final example we turn again to production problems. Consider orders for a product arriving at a plant; each order is concerned with one unit of the product and the mode of arrival is of Poissonian character. Since set-up costs and set-up time for the production line are appreciable, orders are allowed to accumulate until their number reaches a prescribed magnitude. A little reflection leads to the identification of this situation with the model analyzed here.

Both the theory and the examples presented in this paper demonstrate the particularly close affinity of the queueing problems under investigation to storage and inventory theory. It is no coincidence that the optimal R as derived in the preceding Section is represented by a square root formula. This relationship will be developed elsewhere in some detail.

*Transmission by messenger (or any other similar means) is completely equivalent for the purposes of the model.

REFERENCES

1. Benson, F., "Further Notes on the Productivity of Machines Requiring Attention at Random Times", J. Roy. Stat. Soc. B 14, 200-210 (1952).

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

3. Avi-Itzhak, B., Brosh, I. and Naor, P., "On Discretionary Priority Queueing", (to be published) available from University of North Carolina, Institute of Statistics, Mimeo Series No. 338.

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

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

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