In particular,

$$\int_0^\infty x\,dW_1(x) = -B_1^{(1)} - \frac{A_2''}{2A_2'} = \frac{\lambda_1 b_1}{2(1-\lambda_1 a_1)} + \frac{\lambda_2 b_2(1-\lambda_1 a_1)^2 + \lambda_1 b_1(\lambda_2 a_2)^2}{2(1-\lambda_1 a_1)[(1-\lambda_1 a_1)^2(1-\lambda_2 a_2)^2 - (\lambda_1 a_1 \lambda_2 a_2)^2]} \tag{91}$$

and $$\int_0^\infty x^2\,dW_1(x) = B_1^{(2)} + \frac{B_1^{(1)} A_2''}{2A_2'} + \frac{A_2'''}{3A_2'}, \tag{92}$$

where the quantities on the right-hand side are given above.

NOTE

THE QUEUING process discussed in this paper has previously been investigated by B. AVI-ITZHAK, W. L. MAXWELL, AND L. W. MILLER.[1] They used intuitive methods. In this paper a simple and rigorous method is given. Avi-Itzhak, Maxwell, and Miller have found an explicit formula for the expected waiting time. Their formula and formula (91) in the present paper are in agreement. However, it would be interesting to calculate also the higher moments by using both methods and make comparisons.

REFERENCES

1. B. AVI-ITZHAK, W. L. MAXWELL, AND L. W. MILLER, "Queuing with Alternating Priorities," *Opns. Res.* **13**, 306–318 (1965).
2. L. TAKÁCS, "Delay Distributions for One Line with Poisson Input, General Holding Times, and Various Orders of Service," *Bell System Technical J.* 42, 487–503 (1963).

Operations Research 1968 v 16 #3

SOME INEQUALITIES IN QUEUING†

K. T. Marshall

Bell Telephone Laboratories, Inc., Holmdel, New Jersey

(Received April 26, 1967)

Bounds are found for various measures of performance in certain classes of the $GI/G/1$ queue. First, the mean wait in queue is found in terms of the mean and variance of the interarrival, service, and *idle* distributions. Bounds on the idle time moments lead to bounds on the mean wait and number in queue. The interarrival time distribution is then assumed to have mean residual life bounded above by $1/\lambda$ (λ=arrival rate); i.e., given a time t since the last arrival, the expected time to the next arrival is no more than $1/\lambda$. With this assumption the mean number in queue (and hence system) is bounded to within $(1+\rho)/2$ customers. Both upper and lower bounds are tight. The stronger assumption that, given time t since the last arrival, the probability an arrival occurs in the next Δt is nondecreasing in t, leads to bounds on the mean queue length to within $(c_a^2+\rho)/2$, where c_a is the coefficient of variation of the arrival distribution. Again the bounds are tight. Specializing to the $D/G/1$ queue the mean queue length is found to within $\rho/2<\frac{1}{2}$ customer.

LITTLE work has been done on approximations in queuing. Emphasis has been on complex analytic results. Notable exceptions are papers by KINGMAN[2,3] and recently by NEWELL.[6] The paper by Newell is applied primarily to traffic light problems, whereas Kingman's is more closely related to this paper.

PART 1. SOME RESULTS AND BOUNDS FOR ALL $GI/G/1$ QUEUES

Some new results are found for various indicators of performance in the $GI/G/1$ queue. Bounds that are easily calculable are found for such items as the expected wait in queue, expected length of an idle period and the variance of interoutput times.

We find a relation between the idle time between busy periods and the waiting time of a customer in queue. The expected wait in queue is found in terms of the first two moments of the interarrival, service, and idle times. For Poisson arrivals the idle time distribution is exponential, and the expected wait is calculated easily. In general, the moments of the idle distribution are difficult to calculate. However, an upper bound for all

† This paper is based on part of a Ph.D. dissertation submitted to the Department of Industrial Engineering and Operations Research at the University of California, Berkeley, September 1966. This research was supported by the Office of Naval Research under contract Non-r-222(83), and by the Army Research Office, Durham, under contract DA-31-124-ARO-D-331.

$GI/G/1$ queues is easily found in terms of the mean and variance of the arrival and service streams only (*see also* Kingman[2,3]). A lower bound is found that requires knowledge of the arrival and service distributions, and not just the first two moments. In the derivations *FIFO* order of service is assumed, but with the exception of the variance of the wait all results are independent of this assumption.

Notation

We shall deal exclusively with stationary queues in this paper, by which we shall mean that the queuing process either started at time zero with stationary conditions or that it started with some initial condition (such as the wait in queue of the first customer is zero) but that time was at $-\infty$. Hence $W_n \sim W(t)$ for all n.

The following notation is used throughout the paper. The sign $\sim$ is used to signify 'with distribution function.'

The subscript n (e.g., W_n) refers to the nth customer in a stationary stream. When it is not required to note the order of the customers the subscript will be dropped.

T_n = time between nth and $(n+1)$th arrival, $T_n \sim A(t)$, $E[T_n]=1/\lambda$.
S_n = service time of nth customer, $S_n \sim G(t)$, $E[S_n]=1/\mu$.
$U_n = S_n - T_n$, $U_n \sim K(t)$.
τ_n = time between nth and $(n+1)$th departure.
$\rho = \lambda/\mu$.
W_n = wait in queue of nth customer, $W_n \sim W(t)$.
I = length of idle period between busy periods, $I \sim H(t)$.
B = length of busy period, $B \sim B(t)$.

It is possible in some queuing situations that an arrival and service can take place together, leading to problems in defining what is an idle period for the queue. We shall define $P[I=0]=0$, and thus if an arrival occurs at the instant the last customer present departs, the busy period continues, and ends *only* when the facility is empty for a positive length of time.

N_b = number served in a busy period.
D = total delay in system $= W+S$, $D \sim W^*(t)$.
N_q = number in the queue at a random point in time.
$\nu_f^{(n)}$ = nth moment about origin of random variable with distribution F.
The superscript is dropped for $n=1$, e.g., $\nu_a = 1/\lambda$, $\nu_g = 1/\mu$, $\nu_h = E[I]$.
σ_f^2 = variance of a random variable with distribution F.
$c_f^2 = \sigma_f^2/(\nu_f)^2$, where c_f is the coefficient of variation.
$a_0 = P[\text{Arrival finds the system empty}] = P[W_n + U_n < 0]$†
$F^c(t) = 1 - F(t)$ for any distribution F.

† Note that $a_0 = P[W_n = 0]$ only if $P[W_n + U_n = 0]\,\lambda = 0$. This is not necessarily the case in this paper as is pointed out in the discussion of the idle time above.

The Wait in Queue and the Idle Period

Some relations between the moments of the arrival, service, idle, and waiting time distributions are now found. We have the well-known equation

$$W_{n+1} = \max[0, W_n + U_n]. \tag{1}$$

Let $X_n = -\min[0, W_n + U_n]$. Hence

$$W_{n+1} - X_n = W_n + U_n, \tag{2}$$

where $X_n > 0 \Rightarrow X_n = I$. Taking expectations in (2) (with $\rho < 1$) we have

$$a_0 E[I] = 1/\lambda - 1/\mu. \tag{3}$$

This result is given in Rice[8] and Riordan.[7] The result holds for more general queues than the $GI/G/1$, but this fact will not be used in the remainder of this paper.

As examples, for Poisson arrivals $a_0 = (1-\rho)$ and the idle distribution is exponential with mean $1/\lambda$. For the constant arrival, constant service case ($D/D/1$ queue) $a_0 = 1$ and $I = 1/\lambda - 1/\mu$.

An expression is now derived for the expected wait in queue.

Theorem 1. *For all $GI/G/1$ queues with $\rho < 1$,*

$$\begin{aligned} E[W] &= \{E[U^2]/-2E[U]\} - \{E[I^2]/2E[I]\} \\ &= [\lambda^2(\sigma_a^2 + \sigma_g^2) + (1-\rho)^2]/2\lambda(1-\rho) - \nu_h^{(2)}/2\nu_h. \end{aligned} \tag{4}$$

Proof. Square both sides of (2) and note that $W_{n+1}X_n = 0$, giving

$$W_{n+1}^2 + X_n^2 = W_n^2 + 2W_nU_n + U_n^2.$$

Taking expectations, since W_n and U_n are independent,

and $$E[X_n^2] = a_0 E[I^2],$$

we have $$a_0 E[I^2] = 2E[U_n]E[W_n] + E[U_n^2].$$

Using (3) the result follows.

It is interesting to note the special way in which the moments of the idle distribution occur. $[\nu_h^{(2)}/2\nu_h]$ is the mean of an equilibrium excess idle distribution, that is, it is the mean of a random variable with distribution function

$$\int_0^t \frac{H^c(u)}{\nu_h}\,du.$$

This is a well-known result in renewal theory (*see*, for example, reference 1).

Consider again our two examples. For Poisson arrivals $\nu_h^{(2)}/2\nu_h = 1/\lambda$. In this case (4) reduces to

$$E[W]=\rho(1+c_g^2)/2\mu(1-\rho),$$

a well-known result. For the $D/D/1$ queue $\sigma_a^2=\sigma_g^2=0$ and

$$\nu_h^{(2)}=(\nu_h)^2=(1/\lambda-1/\mu)^2,$$

in which case (4) reduces to $E[W]=0$.

An expression for the variance of the wait is now found in a similar manner and is given by

THEOREM 2. *For all $GI/G/1$ queues with $\rho<1$, and FIFO order of service,*

$$\sigma_w^2=[E[U^3]/-3E[U]]+[E[U^2]/-2E[U]]^2+[\{E[I^3]/3E[I]\}-E[I^2]/2E[I]^2],$$

or

$$\sigma_w^2=[\lambda(\nu_g^{(3)}-\nu_a^{(3)})+3(\rho\nu_a^{(2)}-\nu_g^{(2)})]/3(1-\rho)$$
$$+[\{\lambda^2(\sigma_a^2+\sigma_g^2)+(1-\rho)^2\}/2\lambda(1-\rho)]^2+\sigma_{e,h}^2, \qquad (5)$$

where

$$\sigma_{e,h}^2=\nu_h^{(3)}/3\nu_h-[\nu_h^{(2)}/2\nu_h]^2.$$

Proof. Write (2) as $W_{n+1}-X_n=W_n+U_n$ and cube both sides. Note that $X_n^2W_{n+1}=W_{n+1}^2X_n=0$. Using (3) and (4) after taking expectations the result follows.

It has been assumed in the above proofs that the necessary moments exist. Equation (18) in Part 2 shows that a sufficient condition for this to be true is that the first three moments of $A(t)$ and $G(t)$ exist.

The expression for the expected wait is of particular interest in queuing and it is seen to depend only on the first two moments of the interarrival, service, and idle distributions. In general these idle period moments are difficult to calculate but bounds will be obtained for them in later sections of this paper.

The Variance of the Output

It is obvious that $E[\tau_n]=E[T_n]=1/\lambda$. The variance is found as follows. Since $\tau_n=S_{n+1}+X_n$, S_{n+1}, X_n independent,

$$\mathrm{var}[\tau_n]=\mathrm{var}[S_{n+1}]+\mathrm{var}[X_n]. \qquad (6)$$

From (2)

$$\mathrm{var}[W_{n+1}-X_n]=\mathrm{var}[D_n-T_n]=\sigma_a^2+\sigma_g^2+\sigma_w^2. \qquad (7)$$

But

$$\mathrm{var}[W_{n+1}-X_n]=\sigma_w^2+\mathrm{var}[X]-2\mathrm{cov}(W_{n+1}X_n). \qquad (8)$$

Now $W_{n+1}X_n=0$ and hence,

$$\mathrm{cov}(W_{n+1}X_n)=-E[W](1/\lambda-1/\mu).$$

Using this with (6), (7), and (8) gives

$$\mathrm{var}[\tau_n]=\sigma_a^2+2\sigma_g^2-(2/\lambda)(1-\rho)E[W]. \qquad (9)$$

Using equation (4) for $E[W]$ we have finally,

$$\mathrm{var}(\tau_n)=\sigma_g^2-[(1-\rho)^2/2]+[(1-\rho)/\lambda](\nu_h^{(2)}/\nu_h). \qquad (10)$$

For the $M/G/1$ queue this gives $\mathrm{var}(\tau_n)=\sigma_g+[(1-\rho^2)/\lambda^2]$, so that the variance of the output of the $M/G/1$ queue is known exactly when the mean and variance of the service distribution are given.

If $G(t)$ is exponential, $\sigma_g^2=1/\mu^2$ and $\mathrm{var}(\tau_n)=1/\lambda^2$.

In the case of constant arrivals, constant service, $\sigma_g^2=0$ and $\mathrm{var}(\tau_n)=0$.

Some Bounds for All GI/G/1 Queues

Using the results of the previous sections, some simple bounds can be found for various factors in the $GI/G/1$ queue, such as the mean length of an idle period and the mean wait in queue.

(a) *The mean idle time.* Since $a_0\leqq 1$, (3) immediately gives a lower bound on the mean length of an idle period,

$$E[I]\geqq(1/\lambda)-(1/\mu). \qquad (11)$$

The bound is tight for the $D/D/1$ queue.

(b) *The wait in queue.* From equation (4) using (11) and $\mathrm{var}[I]\geqq 0$, it follows that

$$E[W]\leqq\lambda(\sigma_a^2+\sigma_g^2)/2(1-\rho). \qquad (12)$$

This upper bound for all $GI/G/1$ queues is also derived by Kingman.[2,3] Equality holds for the $D/D/1$ queue.

The importance of these bounds is that they involve at most only the first two moments of the arrival and service distributions and further knowledge of the distributions is not required. However, if $K(t)$ is known (or alternatively if $A(t)$ and $G(t)$ are known) a lower bound on the wait in queue can be found as follows.

THEOREM 3. *Let l be a solution of*

$$x=\int_{-x}^{\infty}K^c(u)\,du, \qquad x\geqq 0, \qquad where \qquad (S_n-T_n)\sim K(t)$$

which exists and is unique if and only if $\rho<1$. Then for all GI/G/1 queues, $E[W]\geqq l$.

Proof. Recall the fundamental equation (1)

$$W_{n+1}=\max[0,\ W_n+U_n].$$

Then

$$[W_{n+1}|W_n=x]=\max[0,\ x+U_n],$$

and
$$E[W_{n+1}|W_n=x]=\int_{-x}^{\infty}K^c(u)\,du \qquad \text{all} \qquad x\geqq 0. \qquad (13)$$

Now let
$$\int_{-x}^{\infty}K^c(u)\,du=g(x),$$

which is a continuous convex function for $x\geqq 0$, with $g'(x)=K^c(-x)$, so $K^c(0^+)=g'(0^-)=P[U_n>0]$ and $g'(x)\to 1$ as $x\to\infty$. Let
$$-\beta=E[\min(0,\,U_n)]=\int_{-\infty}^{0}K(u)\,du$$

and
$$\alpha=E[\max(0,\,U_n)]=\int_{0}^{\infty}K^c(u)\,du.$$

Then
$$\alpha-\beta=(1/\mu)-(1/\lambda).$$

From (13)
$$E[W_{n+1}]=\int_{0^-}^{\infty}g(x)\,dW_n(x),$$

or
$$E[W_{n+1}]=E[g(W_n)].$$

Using Jensen's inequality for the expected value of a convex function of a nonnegative random variable,
$$E[W_{n+1}]\geqq g(E[W_n]),$$

so that
$$E[W]\geqq\int_{-E[W]}^{\infty}K^c(u)\,du. \qquad (13a)$$

Consider the equation
$$x=\int_{-x}^{\infty}K^c(u)\,du, \qquad (x\geqq 0) \qquad (14)$$

This can be written
$$x=\alpha+\int_{-x}^{0}K^c(u)\,du. \qquad (x\geqq 0)$$

The situation is drawn in Fig. 1. The equation has a solution if and onl if the two curves cross. If $\alpha=0$, $x=0$ is a solution; if $\alpha>0$ the curves cro if and only if for x sufficiently large,
$$x>\alpha+\int_{-x}^{0}K^c(u)\,du\Leftrightarrow\int_{-x}^{0}K(u)\,du>\alpha,$$

or if and only if $\beta>\alpha$. But $\beta>\alpha$ if and only if $1/\lambda>1/\mu$. Uniquenes comes from convexity arguments. Uniqueness fails only when the tw curves coincide over some range, $[a, b]$ say. This implies $g'(x)=K^c(-x)=$

1 on $[a, b]\Rightarrow g'(x)=1$ on $[a, \infty)\Rightarrow$curves do not cross. In the case $\rho\geqq 1$, either no solution exists, or, for example in the case of the $D/D/1$ queue, an infinite number of solutions exist with $\rho=1$.

So for $\rho<1$, let l be the unique solution of (14). It is now shown that

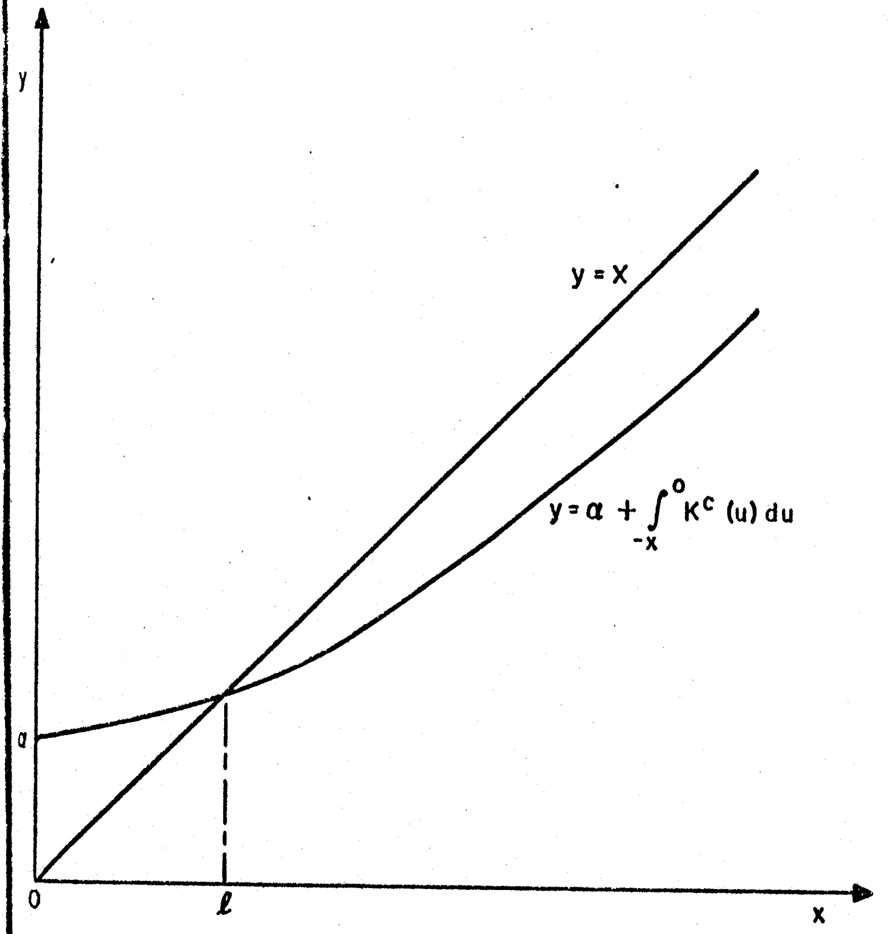

Fig. 1. Determination of the lower bound on the wait in queue.

$l\leqq E[W]$. This is obvious from Fig. 1 and equations (13) and (14). If $l=0$ the inequality is trivial. If $l>0$, then $\alpha>0$ and for all $0\leqq x<l$,
$$x<\alpha+\int_{-x}^{0}K^c(u)\,du$$

from the uniqueness property of l. Hence, if $E[W]<l$, then

$$E[W]<\int_{-E[W]}^{\infty} K^c(u)\,du,$$

which contradicts (13a) and the theorem is proved.

Summarizing, we have shown that for all $GI/G/1$ queues with $\rho<1$

$$l\leqq E[W]\leqq\lambda(\sigma_a^2+\sigma_g^2)/2(1-\rho), \qquad (15)$$

where l is the unique solution of (14). For $\sigma_a^2+\sigma_g^2>0$ (i.e., all except the $D/D/1$ queue), both bounds tend to infinity as $1/\lambda\to 1/\mu>0$. However, their ratio may diverge in a particular case as is shown below for the case of the $M/M/1$ queue.

For the Poisson arrival, exponential service queue it is found that

$$k(t)=(\lambda\mu/\mu+\lambda)e^{-\mu t} \qquad (t\geqq 0)$$
$$=(\lambda\mu/\mu+\lambda)e^{\lambda t}, \qquad (t\leqq 0)$$

which gives

$$K^c(t)=[\rho/(1+\rho)]e^{-\mu t} \qquad (t\geqq 0)$$
$$=1-[1/(1+\rho)]e^{\lambda t}. \qquad (t\leqq 0)$$

Using this in (14) it is found that the lower bound for this case is given by:

$$l=-(1/\lambda)\log_e(1-\rho^2), \qquad \text{which}\to\infty \text{ as } \rho\to 1^-.$$

However, it is easy to show that $\lim_{\rho\to 1^-}(1-\rho)\log_e[1/(1-\rho^2)]=0$ and hence, the bounds diverge. The upper and lower bounds and true value of $E[W]$ are shown in Fig. 2 for fixed $\lambda=1$ and varying μ.

(c) *The variance of the output.* The variance of the output distribution is given in equation (9). Using arguments similar to those in (b) the following upper and lower bounds are found for all general arrival, general service single channel queues,

$$\sigma_g^2\leqq \text{var}[\tau_n]\leqq\sigma_a^2+2\sigma_g^2-2l(1/\lambda-1/\mu), \qquad (16)$$

where l is the solution of equation (14).

PART 2. BOUNDS FOR TWO SUBCLASSES OF $GI/G/1$ QUEUES

IN PART 1 it was seen that the moments of the idle time distribution occurred in many of the expressions. The idle time distribution is some complicated tail distribution of an interarrival time and it might be conjectured that by placing some restriction on the interarrival time distribution one might obtain some desirable properties of the moments of the idle period regardless of the service distribution. This indeed turns out to be true.

Three restrictions on $A(t)$ will be applied in turn in increasing order of strength. In the following definitions and in the remainder of this paper the words 'decreasing' and 'increasing' are used in the weak sense. They should always be read to mean nonincreasing and nondecreasing respectively, and will be given the symbols ↓ and ↑. Expressions, symbols, and words in parentheses should be read together.

320

Definition 1. A nondiscrete distribution F has its mean residual life bounded above (below) by γ, denoted γ-*MRLA* (γ-*MRLB*) if and only if

$$\int_t^\infty \frac{F^c(u)\,du}{F^c(t)} \underset{(\geqq)}{\leqq} \gamma \qquad \text{all} \quad t\geqq 0, \quad \text{where} \quad \gamma<\infty.$$

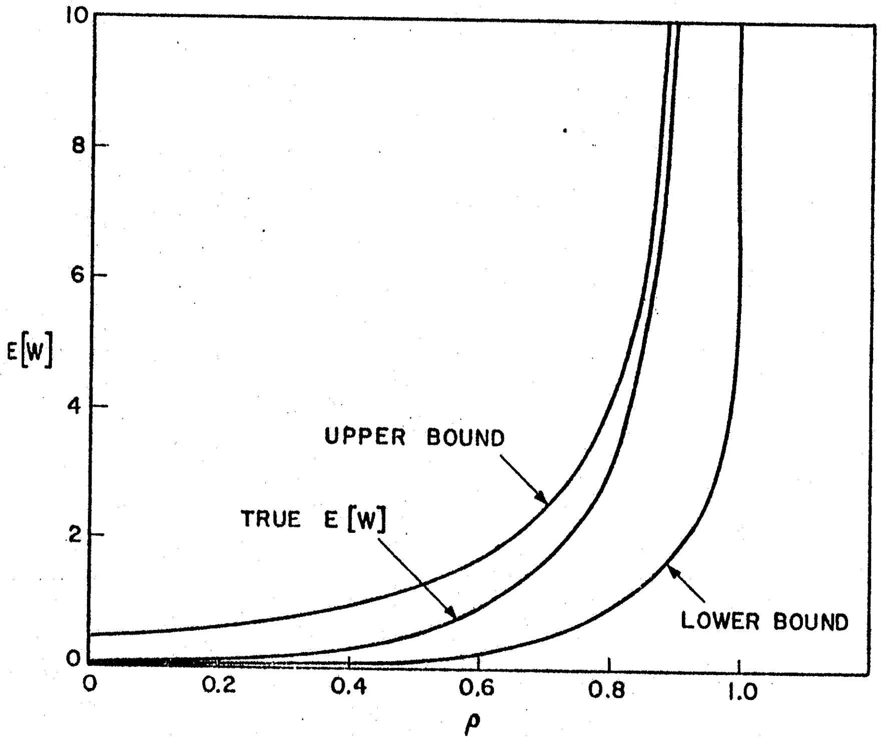

Fig. 2. Bounds on the expected wait in the $M/M/1$ queue.

Definition 2. A nondiscrete distribution F has decreasing (increasing) mean residual life, denoted *DMRL* (*IMRL*), if and only if

$$\int_t^\infty \frac{F^c(u)\,du}{F^c(t)} \underset{(\uparrow)}{\downarrow} \qquad \text{for all} \quad t\geqq 0 \quad F^c(t)>0.$$

Definition 3. A nondiscrete distribution F has increasing (decreasing) failure rate, denoted *IFR* (*DFR*), if and only if, for any $\Delta>0$,

$$[F(t+\Delta)-F(t)]/F^c(t)\uparrow(\downarrow) \quad \text{for all} \quad t\geqq 0 \quad \text{where} \quad F^c(t)>0.$$

Definitions corresponding to these can be given for F discrete, but to simplify the notation and avoid repetition we shall usually assume that F is nondiscrete. We always assume that $F(0^-)=0$.

These concepts are widely used in reliability theory where strong physical justifications can be given for their use in particular problems. In queuing an *IFR* arrival distribution would have the following physical interpretation. Given it has been a time t since the last customer arrived, the probability that a customer arrives in the next small interval Δ is increasing in t. Besides any physical justification many parametric families have this property; for example the gamma and Weibul distributions in certain parameter ranges, and the truncated normal and modified extreme value distributions. The degenerate distribution of the constant arrival queue also has the *IFR* property. It is easy to show that for $F(t)$,

$$IFR(DFR) \Rightarrow DMRL(IMRL) \Rightarrow \nu_f - MRLA\,(\nu_f - MRLB).$$

For a fuller discussion on these properties the reader should consult Chap. 2 of R. E. Barlow and F. Proschan.[1]

For $\nu_a - MRLA/G/1$ queues, (that is, the class of $GI/G/1$ queues whose arrival distributions have the $\nu_a - MRLA$ property) it is shown that simple expressions can be obtained to bound, for example, the expected number in the queue to within *at most* one customer. These bounds involve only the mean and variance of the arrival and service streams. For the special class of $D/G/1$, (constant arrival, general service), the expected number in the queue is bound to within at most one half.

Some Properties of the Idle Time Distribution

In this section three theorems are proved that give some useful properties of the idle distribution. These are used in the next section to bound certain measures of performance in various classes of queues.

Theorem 4. *For the class of $GI/G/1$ queues where $A(t)$ has γ-MRLA (γ-MRLB), denoted γ-MRLA/G/1 (γ-MRLB/G/1) queues,*

$$\nu_h^{(2)}/2\nu_h \leqq (\geqq)\gamma. \tag{17}$$

Equality holds when $A(t)$ is exponential and $\gamma=1/\lambda$.

Before proving this theorem we relate the distributions of idle time and interarrival time, as these relations play a key role in the proofs of all three theorems in this section. Throughout this section the integrals are Lebesque-Stieltjes integrals.

In the third section of Part 1 we defined $X_n=-\min[0, W_n+U_n]$, and noticed that if $X_n>0$ then $X_n=I$. If $X_n=0$ an idle period does not occur

321

after customer n. Hence I is only defined on that part of the sample space of the sequence $\{X_n\}$ for which the X_n take on positive values.

For any $t\geqq 0$

$$P[X_n>t]=P[T_n-(W_n+S_n)>t]$$
$$=\int_0^\infty A^c(t+x)\,dW_n{}^*(x).$$

Since we are assuming stationarity,

$$H^c(t)=P[X>t|X>0]=P[X>t]/P[x>0],$$

or

$$H^c(t)=M\int_0^\infty A^c(t+x)\,dW^*(x), \tag{18}$$

where the normalizing constant†

$$M^{-1}=\int_0^\infty A^c(y)\,dW^*(y)=a_0.$$

Proof of Theorem 4. Using (18) we have

$$\int_t^\infty H^c(u)\,du=M\int_t^\infty\int_0^\infty A^c(u+x)\,dW^*(x)\cdot du$$
$$=M\int_0^\infty\int_t^\infty A^c(u+x)\,du\cdot dW^*(x),$$

since the integrals converge absolutely. Making a change of variable and multiplying numerator and denominator of the right-hand side by $A^c(t+x)$ we get

$$\int_t^\infty H^c(u)\,du=M\int_0^\infty A^c(t+x)\int_{t+x}^\infty \frac{A^c(v)}{A^c(t+x)}\,dv\cdot dW^*(x)$$
$$\overset{\leqq}{(\geqq)}\gamma M\int_0^\infty A^c(t+x)\,dW^*(x)$$

from the γ-MRLA (γ-MRLB) assumption. Hence

† The author is indebted to William S. Jewell for this approach. Prepublication discussions with the referees has lead to a clearer presentation of the derivation of (18) for which the author is grateful. In reference 5 a different approach is taken and a different representation of $H(t)$ is obtained, namely $H^c(t)=\int_0^\infty \frac{A^c(t+x)}{A^c(x)}\,d\Phi(x)$ where $\Phi(t)$ is the distribution of total delay of the *last* customer in a busy period. The author is indebted to one of the referees for showing the equivalence of these representations, which, it is understood, will appear together with more discussion of (18) in a letter to the editor.

$$\int_t^\infty \frac{H^c(u)\,du}{\nu_h} \overset{\leq}{(\geq)} \frac{\gamma H^c(t)}{\nu_h} \qquad \text{all} \qquad t \geq 0, \tag{19}$$

and integrating over t on both sides proves the theorem.

THEOREM 5. *For the class of GI/G/1 queues where $A(t)$ has DMRL (IMRL), denoted DMRL/G/1 (IMRL/G/1) queues,*

$$\int_t^\infty \frac{H^c(x)\,dx}{H^c(t)} \overset{\leq}{(\geq)} \int_t^\infty \frac{A^c(x)\,dx}{A^c(t)} \qquad \text{all} \qquad t \geq 0.$$

Equality holds when $A(t)$ is exponential.

Proof. The proof is essentially the same as that in Theorem 4, but applying the *DMRL* (*IMRL*) assumption. Details are left to the reader.

THEOREM 6. *For the class of GI/G/1 queues where $A(t)$ has IFR (DFR), denoted IFR/G/1 (DFR/G/1) queues,*

(i) $[H(t+\Delta)-H(t)]/H^c(t) \geq (\leq) [A(t+\Delta)-A(t)]/A^c(t)$ *for $\Delta > 0$, and all $t \geq 0$ where finite,*

(ii) $H^c(t)/A^c(t) \downarrow (\uparrow)$ all $t \geq 0$,

(iii) $$\int_t^\infty \frac{H^c(u)\,du}{\nu_h} \overset{\leq}{(\geq)} \int_t^\infty \frac{A^c(u)\,du}{\nu_a} \qquad \text{all} \qquad t \geq 0.$$

When $A(t)$ is exponential (i) and (iii) are equalities and the ratio (ii) is equal to 1.

Proof. (i) Using (18), since $H(t+\Delta)-H(t) = H^c(t) - H^c(t+\Delta)$,

$$\frac{H(t+\Delta)-H(t)}{H^c(t)} = \frac{M}{H^c(t)} \int_0^\infty [A^c(t+x) - A^c(t+x+\Delta)]\,dW^*(x)$$
$$= \frac{M}{H^c(t)} \int_0^\infty \frac{[A(t+x+\Delta)-A(t+x)]}{A^c(t+x)} A^c(t+x)\,dW^*(x)$$
$$\overset{\geq}{(\leq)} \frac{A(t+\Delta)-A(t)}{A^c(t)}$$

from the *IFR* (*DFR*) assumption.

(ii) Add and subtract 1 from both sides of (i),

$$1 - \frac{H^c(t+\Delta)}{H^c(t)} \overset{\geq}{(\leq)} 1 - \frac{A^c(t+\Delta)}{A^c(t)},$$

or

$$\frac{H^c(t)}{A^c(t)} \overset{\geq}{(\leq)} \frac{H^c(t+\Delta)}{A^c(t+\Delta)} \qquad \text{all} \qquad \Delta > 0, \quad t \geq 0,$$

which proves part (ii). Notice that (i) and (ii) are equivalent as this argument is reversible.

(iii) From Theorem 5

$$\int_t^\infty H^c(u)\,du \Big/ \int_t^\infty A^c(u)\,du \overset{\leq}{(\geq)} \frac{H^c(t)}{A^c(t)} \overset{\leq}{(\geq)} \frac{H^c(v)}{A^c(v)} \qquad \text{all} \qquad 0 \leq v \leq t$$

from (ii) and the fact that *IFR*(*DFR*)⇒*DMRL*(*IMRL*) (*see* Barlow and Proschan[11]). Putting this in determinant form,

$$\begin{vmatrix} \int_t^\infty H^c(u)\,du & H^c(v) \\ \int_t^\infty A^c(u)\,du & A^c(v) \end{vmatrix} \overset{\leq}{(\geq)} 0 \qquad \text{all} \qquad 0 \leq v \leq t.$$

Integrating v over $(0, t)$

$$\begin{vmatrix} \int_t^\infty H^c(u)\,du & \int_0^t H^c(u)\,du \\ \int_t^\infty A^c(u)\,du & \int_0^t A^c(u)\,du \end{vmatrix} \overset{\leq}{(\geq)} 0.$$

Adding the first column to the second gives

$$\begin{vmatrix} \int_t^\infty H^c(u)\,du & \nu_h \\ \int_t^\infty A^c(u)\,du & \nu_a \end{vmatrix} \overset{\leq}{(\geq)} 0,$$

which proves (iii) and completes the proof of the theorem.

Part (iii) leads to the following

COROLLARY. *For IFR/G/1 queues (DFR/G/1 queues)*

$$\frac{\nu_h^{(2)}}{2\nu_h} \overset{\leq}{(\geq)} \frac{\nu_a^{(2)}}{2\nu_a} = \frac{\lambda}{2}\left[\sigma_a{}^2 + \left(\frac{1}{\lambda}\right)^2\right] = \frac{(c_a{}^2+1)}{2\lambda}. \tag{20}$$

Equality is taken on by the M/G/1 queue.

Bounds for Two Subclasses of the GI/G/1 Queue

(a) *The mean idle time and the probability an arrival finds the system empty.*

Recall from Part 1, equation (3) that

$$a_0\nu_h = (1/\lambda)(1-\rho).$$

Using (19) above and (11) in Part 1 gives

(i) For $1/\lambda$-*MRLA*/*G*/1 queues,

$$(1-\rho) \leq a_0 \leq 1,$$
$$(1/\lambda)(1-\rho) \leq \nu_h \leq 1/\lambda.$$

(ii) For $1/\lambda$-*MRLB/G*/1 queues,

$$0 \leqq a_0 \leqq (1-\rho),$$

$$1/\lambda \leqq \nu_h.$$

The upper bound in (i) and lower bound in (ii) are taken on by the Poisson arrival queue. The lower bound in (i) is taken on by the $D/D/1$ queue.

From the relations

$$E[B] = [\rho/(1-\rho)]E[I] \quad \text{and} \quad E[N_b] = \mu E[B]$$

one can obtain simple bounds on the mean length of and number served in a busy period.

(b) *The mean wait and number in queue.* Using (4) and (12) from Part 1 with Theorem 4 gives

(i) For all $1/\lambda$-*MRLA/G*/1 queues with $\rho<1$,

$$J-[(1+\rho)/2\lambda] \leqq E[W] \leqq J, \tag{21}$$

$$\lambda J-[(1+\rho)/2] \leqq E[N_q] \leqq \lambda, \tag{22}$$

where

$$J = (c_a^2+\lambda^2\sigma_g^2)/2\lambda(1-\rho). \tag{23}$$

Equation (22) follows from (21) by applying the important queuing formula $E[N_q] = \lambda E[W]$ (*see* Little.[4])

Equation (21) shows that for a broad class of queues the mean wait in queue (or system) has been bound to within *at most* a mean interarrival time. Equation (22) gives bounds on the expected number in queue that differ by *at most one* customer. The lower bounds are taken on by the $M/G/1$ queue, the upper ones by the $D/D/1$ queue.

These bounds are linear in the variance of the arrivals and the variance of the service, for any fixed ρ. The narrowness of the bounds would suggest that the mean wait and mean number in queue increase 'approximately linearly' with these variances.

By making the stronger assumption of *IFR* arrivals, using (4) and (12) in Part 1 with the corollary to Theorem 6 we obtain

(ii) For all *IFR/G*/1 queues with $\rho<1$,

$$J-(c_a^2+\rho)/2\lambda \leqq E[W] \leqq J, \tag{21a}$$

$$\lambda J-(c_a^2+\rho)/2 \leqq E[N_q] \leqq \lambda J, \tag{21b}$$

where J is given in (23). Again there is equality with the lower bounds with the $M/G/1$ queue, and with the upper bounds with the trivial $D/D/1$ queue. A property of *IFR* arrivals is that $c_a^2 \leqq 1$ (*see* Barlow and Proschan[1]).

An important special subclass of these queues are those with constant interarrival times (the $D/G/1$ queue). In these queues $c_a^2=0$ so that (21b) gives the mean number in queue to within *at most* ½ customer.

Using the results obtained so far bounds for the other subclasses of the $GI/G/1$ queue are easily obtained. For example, using (15) and (20), for all *DFR/G*/1 queues with $\rho<1$

$$l \leqq E[W] \leqq J-(c_a^2+\rho)/2\lambda.$$

As shown in Part 1 these bounds may diverge. In the interests of brevity, since no new techniques are involved, the bounds for each class will not be written out explicitly.

It should be noted that bounds in terms of the first moments of $A(t)$ and $G(t)$ only, such as those for a_0 and $E[I]$, are not improved by making the *DMRL* or *IFR* assumption in place of the $1/\lambda$-*MRLA* assumption. In the case of $E[N_q]$ or $E[W]$ the *DMRL* assumption gives no improvement over the bounds obtained under $1/\lambda-$*MRLA*.

Using the results obtained so far it is easy to obtain bounds on such quantities as the variance of the wait and the variance of the output under the various assumptions on $A(t)$. Many of these are given explicitly in Marshall.[5]

ACKNOWLEDGMENTS

THE AUTHOR would like to express his appreciation for the many helpful comments and suggestions of RONALD W. WOLFF whose interest stimulated much of this paper. Acknowledgments are also made to WILLIAM S. JEWELL and GORDON F. NEWELL for their continued interest and support.

REFERENCES

1. R. E. BARLOW AND F. PROSCHAN, *Mathematical Theory of Reliability*, Wiley, New York, 1965.
2. J. F. C. KINGMAN, "On Queues in Heavy Traffic," *J. Roy. Stat. Soc. B* 24, 383–392 (1962).
3. ———, "Some Inequalities for the *GI/G*/1 Queue," *Biometrika* 49, 315–324 (1962).
4. J. D. C. LITTLE, "A Proof of the Queuing Formula $L = \lambda W$," *Opns. Res.* 9, 383–387 (1961).
5. K. T. MARSHALL, "Some Inequalities for Single Server Queues," PhD. Thesis, Dept. of Industrial Engineering, University of California, Berkeley, 1966.
6. G. F. NEWELL, "Approximation Methods for Queues With Application to the Fixed Cycle Traffic Light," *SIAM Rev.* 7, 223–240 (1965).
7. J. RIORDAN, *Stochastic Service Systems*, pp. 77–78, Wiley, New York, 1962.
8. S. O. RICE, "Single Server Systems—I. Relations Between Some Averages," *Bell Sys. Tech. J.* 41, 269–278 (1962).

COMMENTS ON "SOME INEQUALITIES IN QUEUING" BY K. T. MARSHALL

Richard V. Evans

Case Western Reserve University, Cleveland, Ohio

(Received December 15, 1967)

IN THE prepublication discussions of MARSHALL's paper the validity of his equation (18) was questioned. This question was confounded further by the comment in the footnote following the equation that an alternative discussion could be given. This note is intended to elaborate on these questions and hopefully to help any reader who also is unsure about these questions and their relations.

To start it might be well to consider the following random variables although they represent an expansion of the author's system.

I_n' = length of the idle period between nth and $n+1$ customer ($I_n'=0$ if there is no idle time between customers).

I_m = length of the mth idle period in the system.

Moreover, let $I=\lim_{m\to\infty} I_m$ assuming that it exists. It is the distribution of I that is involved in equation (18). The correspondence between the sequences $\{I_m\}$ and $\{I_n'\}$ is of course that $I_m=I'_{n(m)}$, i.e., the mth idle period is the idle time between two customers $n(m)$ and $n(m)+1$ and conversely if $I_n'>0$, then it is one of the sequences $\{I_m\}$. The subsequences of $\{I_n'\}$ of those having positive values is precisely the sequence $\{I_m\}$. The common limit of these subsequences is I. Now the distribution H_n' of I_n' is easily computed in the terms used in the paper.

$$H_n'^c(x)=\int_0^\infty A^c(t+x)\,dW_n^*(t). \qquad (1)$$

This is because the event that $[I_n'>x]=\bigcup_t\{[T_n>t+x]\cap[D_n\simeq t]\}$, where D_n is the time customer n spends in the system, $D_n=W_n+S_n$. Thus, the second event in the intersection is that D_n be approximately t, and the first event in the intersection is that the interarrival time between customers n and $n+1$. These two events are clearly independent and thus (1) follows. Now I_n' is a member of the sequence I_m if $I_n'>0$ and the distribution of I_n' given that it is positive will be given by

$$H_m^c(x)=M_m\int_0^\infty A^c(t+x)\,dW^*_{n(m)}(t),$$

where $M_m^{-1}=\int_0^\infty A^c(t)\,dW^*_{n(m)}(t)$.

For any sequence $n(m)$ the corresponding sequence H_m^c converges to the limit $H^c(x)$ given by

$$H^c(x)=M\int_0^\infty A^c(t+x)\,dW^*(t),$$

$$M^{-1}=\int_0^\infty A^c(t)\,dW(t).$$

Since this is true for any given placement of the idle periods, it is true regardless of such placement, which is Marshall's contention in his (18).

If one approaches the idle times $\{I_m\}$ directly, then I_m corresponds to $I'_{n(m)}$ where the latter is known to be positive. Thus one analyzes the situation in a conditional sample space giving an event relation of the form

$$[I_m>t]=\bigcup{}_x\{[T_{n(m)}>t+x|T_{n(m)}>x]\cap[D_{n(m)}\simeq x|T_{n(m)}>x]\}.$$

Thus

$$H_m^c(t)=\int_0^\infty\frac{A^c(t+x)}{A^c(x)}\,d\varphi_{n(m)}(x),$$

which converges to the relation of the footnote. Now the two are equivalent if one can show that

$$dW^*(x)\Big/\int_0^\infty A^c(x)\,dW^*(x)=d(\varphi x)/A^c(x)$$

or

$$A^c(x)\,dW^*(x)=\left[\int_0^\infty A^c(x)\,dW^*(x)\right]d\varphi(x).$$

The left-hand side is

$$\text{prob}\{[T>x]\cap[D\simeq x]\},$$

since the component events are independent. The right is literally

$$\text{prob}\{[T>D]\cap[D\simeq x|T>D]\}.$$

This second joint event is precisely the same as the first of this pair.

This discussion is still highly heuristic in that the limiting operations that abound have all been assumed to behave properly. The indirect approach to the convergence of $\{I_m\}$, although I think justifiable at any level of rigor desired, has a subtlety that is a bit discomforting. A more direct approach is of course feasible. The approach is to develop in greater detail the relation between $\{I_n'\}$ and $\{I_m\}$. Suppose that the mth idle period occurs between customers $n(m)$ and $n(m)+1$. Given this, what is the distribution of I_{m+1}? One obvious approach is to decompose the event $[I_{m+1}>t]$ according to the possible values of $n(m+1)$, which are $n(m)+k$ for $k=1, 2, \cdots$.

Thus

$$[I_{m+1}>t]=\bigcup_{k=1}^{\infty}\bigcup_{x}\{[T_{n(m)+k}>t+x]\cap[D'_{n(m)+k}\simeq x]\},$$

where the basic events are that appropriate arrival times are sufficiently large and that the delay of customer $n(m)+k$ is approximately x and customer $n(m)+k$ is the kth customer in a busy period that began with the arrival of customer $n(m)+1$. This latter restriction is denoted by using the random variable D_n' with distribution function $W_n^{*\prime}$ for delays under these conditions.

$$H_{m+1}^c(t)=\sum_{k=1}^{\infty}\int_0^{\infty}\{A_{n(m)+k}^c(t+x)\ dW_{n(m)+k}^{*\prime}(x)\}$$

$$=\int_0^{\infty}\{A^c(t+x)\sum_{k=1}^{\infty}dW_k^{*\prime}(x)\}.$$

This, of course, assumes legitimate the manipulations of the integrations. The subscript on A^c may be dropped because of the assumption of common interarrival distribution One can also drop the $n(m)$ part of the subscript on $W_{n(m)+k}^*$ thinking in terms of a prototype busy period that starts with the arrival of a first customer. It appears that for $H_m{}^c(t)$ to converge to the solution to Marshall's (18) we must have

$$\textstyle\sum_{k=1}^{k=\infty} dW_k^{*\prime}(x)=d\varphi(x)/A^c(x)=M\ dW^*(x),$$

or considering the first equality

$$\textstyle\sum_{k=1}^{k=\infty} A^c(x)\ dW_k^{*\prime}(x)=d\varphi(x).$$

This is true since this in-event terms is just the decomposition of the event that the last customer in a busy period has delay approximately x according to which customer in a busy period is the last one.

AN ORDERING POLICY FOR REPAIRABLE STOCK ITEMS†

Stephen G. Allen and Donato A. D'Esopo

Stanford Research Institute, Menlo Park, California

(Received June 12, 1967)

When a stock item fails, it is assumed to be repairable with a known positive probability less than one. In this case stock must be replenished on occasion with new supplies. An ordering policy of the familiar reorder point-order quantity type is considered, and expressions developed for expected shortages, inventory, and number of orders per unit of time. Because shortages within a replenishment cycle can decrease because of returns from repair, the derivation of expected shortages is of particular interest.

WE CONSIDER a system in which a number of identical items are in use but subject to failure. We shall assume that the failure of any one item is independent of the status of the others and that the number of items which fail in a unit of time follows a Poisson law with mean D. When an item fails, it immediately enters a repair cycle with probability p from which it emerges in serviceable condition after a fixed repair time R. A failed item is nonrepairable with probability $1-p$ and is discarded. A stock of serviceable items is normally maintained to replace failed items. We shall finally assume that when serviceable stock is zero and a failure occurs, a backorder is created.

In this paper, we shall study a replenishment policy of the familiar type:

When the total inventory of serviceable items plus items in repair less backorders is reduced to a reorder point X, a replenishment order is immediately placed for Q units that are then received after a fixed lead time L.

Above and in the sequel, the number of items in inventory shall not include those currently in use in the system nor failed items that are nonrepairable.

Our main task will be to derive an expression for the total cost of such a policy, namely, the sum of the expected shortage cost, inventory holding cost, and ordering cost, all per unit time. These three components of total cost are assumed, respectively, to be proportionate to expected units short, inventory, and orders per unit time. Since a decision to *repair versus replenish* is not being considered, the expected repair cost and cost of units

† Presented at the Thirty-First Annual Meeting of the OPERATIONS RESEARCH SOCIETY OF AMERICA, New York City, N. Y., May 31, 1967.