THE OUTPUT RATE OF SINGLE SERVER DEVICES
UNDER AN INHOMOGENEOUS POISSON LOAD

by

Micha Hofri

Technical Report #150
May 1979
ABSTRACT

The output rate - the rate of service completions - in a single server facility is defined and evaluated in terms of its steady state probabilities. The load on the server is Poisson, with rate depending on the size of the queue. The various implications of the results are discussed: an empirical finding on the dependence of the output rate on the service distribution is shown to have a simple explanation. Further, the rate is shown to depend critically on the load on the server, and thus is an imperfect representation of the server characteristics.

Key words and Phrases: Output rate; single server; M/G/1; M/G/1/N; Machine interference.
A. INTRODUCTION

A1. The output rate of a server is a quantity for which standard queueing theory had little use, and is thus rarely treated. Analyses of the output process in its entirety have appeared in various contexts, but usually its dependencies with the input process, or its renewal properties were the locus of investigations. See [4] for a detailed review and further references.

A2. The reciprocal of the output rate is called the service function. This is commonly used to obtain statistical estimates of the service required by a customer.

A3. In recent years a new source of interest in these quantities appeared: more and more service facilities have been designed and implemented that in order to be properly analyzed need to be modeled as an interconnected collection of simpler servers. Prime examples are computing systems and communications networks, but any type of activity which involves the migration of an identifiable entity between various stops can be thus modeled (e.g. issuing a passport, or putting out a fire). A successful approach has been that of decomposition [3]. To carry out such an analysis one often needs to characterize the flow of customers (we use this term generically) from one facility to another in terms of their status. Such a (limited) characterization is the subject of this paper.

A4. A more recent type of analysis that relies, in a sense, on the properties of these quantities is Operational Analysis ([5], [6]). It is this latter formalism that perhaps is more closely related to properties of the output flow treated here.
A5: We limit ourselves to single server subsystems, as they are at once the most important case and easier to handle. We also restrict the analysis to situations where the input process of requests can be represented by an inhomogeneous Poisson stream, with the rate only depending on the current queue length; this is done for reasons of tractability, but portions of the paper, especially Section D, pertain to more general uses. The service regime is limited to be simple (i.e. no processor sharing), nonpreemptive, with the selection for service independent of the service requirements. This embraces FIFO, LIFO and various schemes of external priorities.

A6. Section B defines the models under consideration more formally, and derives the main result. Some calculations are relegated to appendices. Section C presents three important special cases: a standard M/G/1 system, and M/G/1/N system - i.e. an M/G/1-like facility with finite waiting room, and the machine interference problem, which in terms of the examples of A3 can represent a computer servicing a finite set of sources, with each having i.i.exp.d. "think time". Section D contains a discussion of some of the implications of the results obtained. It also contains an explication of an empirical observation that provided the original impetus to this investigation.
B. THE SINGLE SERVER WITH INHOMOGENEOUS POISSON INPUT.

B1. Consider a single server service facility. Customers arrive singly according to a Poisson process. When the system holds \( n \) customers (including the serviced one, if any), the input rate is \( \lambda_n \). Each customer requires a service, \( S \), i.i.d. with the distribution \( F(\cdot) \). We shall often assume the existence of a density, \( f(\cdot) \), but this does not seem to bear on the generality of the results. Denote

\[
\begin{align*}
\alpha &= E(S), \quad \sigma^2 = V(S), \quad L(u) = E(e^{-uS}).
\end{align*}
\]

All the involved variables are assumed independent, as well as the so-called 'service rate' which is taken to be constant, at \( \lambda \) (i.e. the time to process a service requirement of size \( s \)) is \( s \)). The service is nonpreemptible and the selection for service mechanism does not depend on the queue. For our purposes we may assume a FIFO order, but no result is contingent upon this choice.

B2. The 'output rate at time \( t \)' of the server is defined as

\[
r(t) = \lim_{\Delta t \to 0} \frac{P(\text{departure occurs in the interval } (t, t+\Delta t))/\Delta t}{t}. \tag{B-2}
\]

Our interest is directed at the output rate as a function of the system occupancy, \( X \). Thus we define

\[
r(n,t) = \lim_{\Delta t \to 0} \frac{P(X(t) = n; \text{ a departure occurs during } (t, t+\Delta t))}{\Delta t P(X(t) = n)}. \tag{B-3}
\]

Only stable systems are considered in the following analysis. In such systems all the variables which describe the system possess limiting
In particular

\[ P(X(t) = n) = p(n); \quad r(n,t) = r(n); \quad t \to \infty. \]  \hspace{1cm} (B-4)

We wish to calculate \( r(n) \). To this end we need an 'Markovisation' of the system, i.e., an underlying Markov process that encapsulates the information required to calculate probabilities of further evolution.

**B3.** The choice of a sufficient Markovian description is not unique. The one most suitable to our needs specifies the following two variables:

- \( X \) — Number of customers in the system,
- \( V \) — Remaining duration of current service.

For \( X = 0 \) only \( V = 0 \) is used. The limiting joint distribution of \( X \) and \( V \) is

\[ P(n,v) = \text{Prob} \{ \text{the system holds } n \text{ customers, and the remaining portion of the current service requirement is } \leq v \}. \]  \hspace{1cm} (B-6)

This is the probability that a random observation \( \{ r_0 \} \) finds the system in a state \((n,v)\).

For any service distribution, \( P(n,v) \) is absolutely continuous in the second variable, because the distribution of the interarrival period is. Hence \( P(n,v) \) has a density \( p(n,v) \), which we shall mainly use.

A simpler description results if the state of the system is only considered at service termination epochs. Then we talk of the number of customers left behind by a departing customer. Denoting this variable by \( X_d \) we define its limiting pmf

\[ \pi_i = P(X_d = i). \]  \hspace{1cm} (B-7)
B4. Before proceeding to describe the output rate via these variables we establish two equations which will be our main tools in the analysis. The first relates \( p(n,v) \) and \( \pi_i \) by enumerating the sequences of events that may lead to the state \( (n,v) \). An intermediate variable needed is \( D \), with the limiting pmf \( d_i \) is defined as the number of customers at the system in the most recent departure epoch prior to \( t \). Simple renewal arguments show that for \( i > 0 \) \( d_i \) must be proportional to \( \pi_i \), but \( D = 0 \) is the value obtained by observations made during an idle time and the first service period of a busy period. Hence, carrying the required normalization through,

\[
d_i = \left\{ \begin{array}{ll}
\frac{\alpha \lambda_o}{\pi_o + \lambda_o \alpha} & \text{if } i > 0 \\
\frac{1 + \lambda_o \alpha}{\pi_o + \lambda_o \alpha} & \text{if } i = 0
\end{array} \right.
\]  \hspace{1cm} (B-8)

B5. Now randomizing on \( D \) and the service requirement during which the observation took place, \( Y \), we obtain

\[
p(n,v) = \sum_{i=0}^{\infty} \int_{y=v}^{\infty} p(n,v \mid i,y)p(D = i, Y \leq y) \, dy.
\]  \hspace{1cm} (B-9)

In Appendix A this relation is evaluated in terms of the system parameters and we obtain

\[
p(n,v) = \theta \sum_{i=1}^{\infty} \left( \pi_i + \delta_i, \pi_0 \right) \int_{u=0}^{\infty} f(u+v)p(n\mid i,u) \, du
\]  \hspace{1cm} (B-10)

with \( \theta = \lambda_0 / (\pi_0 + \lambda_0 \alpha) \).
The balance equations for \( p(n,v) \) are obtained by considering the evolution in time of these probabilities, from time \( t \) to \( t + \Delta t \). This is done in Appendix B and we obtain

\[
f(v)p(n+1,0^+) = \lambda_n p(n,v) - \lambda_{n-1} p(n-1,v) - \frac{\partial}{\partial v} p(n,v), \quad n \geq 1 \quad (B\text{-}12)
\]

and

\[
p(1,0^+) = \lambda_0 p(0)
\]

where

\[
p(n,0^+) \xrightarrow{v \to 0} \lim p(n,v), \quad p(0,v) \equiv p(0)f(v). \quad (B\text{-}13)
\]

Checking whether \((B\text{-}10)\) satisfy \((B\text{-}12)\) by direct substitution simply yields the balance equations for \( \pi_i \).

\[B7.\text{ The derivation of } (B\text{-}12) \text{ also points to the expression we need for the output rate:}
\]

\[
r(n) = \frac{p(n,0^+)}{p(n)} \quad (B\text{-}14)
\]
where \( p(n) \) is the marginal pmf of \( p(n,v) \). Now the reason for introducing the balance equations (B-12) becomes apparent: they yield a direct estimation of \( p(n+1,0^+) \), in terms of \( p(n,v) \). This is performed in Appendix C, and we get the less apparent result.

\[
\begin{align*}
  p(n + 1,0^+) &= \theta \pi_n, \quad n > 1 \\
  p(2,0^+) &= \theta(\pi_0 + \pi_1) - \lambda_0 p(0),
\end{align*}
\]

with the second part of (B-12) giving \( p(1,0^+) \) directly.

**B8.** Equation (B-15) especially via (B-14) mixes the departure time probabilities \( \pi_i \) and steady state (or "random observation") probabilities \( p(n) \). The two can be related via a simple-minded observation that is detailed in Appendix D, to yield

\[
p(n) = \frac{\theta}{\lambda_n} \pi_n, \quad \theta = \frac{\lambda_0}{(\pi_0 + \lambda_0 \alpha)} \quad n \geq 0, \lambda_n > 0 \tag{B-16}
\]

substituting in (B-15), and using the second part of (B-12) finally gives

\[
p(n+1,0^+) = \lambda_n p(n), \quad n \geq 0 \tag{B-17}
\]

and hence

\[
\begin{align*}
  r(n) &= \lambda_{n-1} p(n-1)/p(n), \quad n \geq 1 \\
  r(n) &= \lambda_n \pi_{n-1}/\pi_n, \quad n \geq 1
\end{align*}
\]

This is the desired result.

**B9.** For reference purposes we also derive now recursion relations for the departure epoch steady state probabilities \( \pi_i \) defined in B3. This
is entirely standard; assuming stationarity:

\[ P(X_d = n) = P(X_d - \delta X_d > 0 + Y X_d = n) \]

(B-19)

where \( Y X_d \) is the number of arrivals during a service period subsequent to a departure which left \( X_d \) in the system.

We note that this variable \( Y_1 \) depends not only on the service duration but also on the occupancy of the system when this duration begins — a standard situation where probabilistic analysis fails to come up with closed form results.

In terms of the notations used in Appendix A, when the value of \( X_d \) is 1, \( P(Y = r) = \text{Prob}(A_1, r - 1 \leq s, A_1, r > s) \), where \( s \) is service duration. Further substitutions from Appendix A yield

\[ \pi_n = \pi_0 \sum_{k=1}^{n} a_k^{(1,n)} \frac{a(n+1)-a(k)}{\lambda_{n+1} - \lambda_k} + \sum_{r=1}^{n} \sum_{k=r}^{\infty} a_k^{(r,n)} \frac{a(n+1)-a(k)}{\lambda_{n+1} - \lambda_k} \]

(B-20)

\[ + \pi_{n+1} a(n+1) \quad n \geq 1 \]

where \( a(k) \) is the value of \( L_s(\cdot) \) at the point \( \lambda_k \). Solving for \( \pi_{n+1} \) and incorporating the equivalent of (B-20) for \( n = 0 \), we obtain

\[ \pi_{n+1} a(n+1) = \pi_n - \delta_0 \pi_0 a(n+1) - \sum_{r=1}^{n} (\pi_{r+\delta} r, 1) \pi_0 \sum_{k=r}^{\infty} a_k^{(r,n)} \frac{a(n+1)-a(k)}{\lambda_{n+1} - \lambda_k} \]

(B-21)

This allows a calculation of all \( \pi_n \) in terms of \( \pi_0 \); for service distribution of phase-type, this should prove a rather painless procedure, although effects of round-off error should be explicitly handled.
C. SPECIALIZATIONS

C1. In this section, three special cases are briefly discussed. The first is a standard M/G/1 queueing facility. This is obtained when \( \lambda_n = \lambda \), for all \( n \geq 0 \). When such is the case, \( X_d \) and \( X \) have the same (limiting) distribution, \( p(n) = \pi_n \). \( \theta \) is simply \( \lambda \) and equations (B-21) have a solution given by the well-known Khinchine-Pollaczek pgf:

\[
G(z) = E(z^X) = (1-p)(1-z)L(z)/(L(z)-z), \quad L(z) = L(\lambda-\lambda z) \quad (C-1)
\]

with \( p = \lambda^\alpha, \pi_0 = 1-p \). It is also possible to thus encapsulate the joint distribution \( p(n,v) \):

\[
H(z,s) = E(z^X e^{-sV}) = (1-p) \left\{ 1 + \lambda z \frac{(1-z)[L(\lambda-\lambda z)-L(s)]}{(s-\lambda+\lambda z)[L(\lambda-\lambda z)-z]} \right\} \quad (C-2)
\]

by directly summing and integrating in (B-10) and using (C-1).

C2. Further then this ability to "close" the solutions for the state probabilities equations there seems to be no special properties of M/G/1 in the present context. We only remark that it is this case which is most suitable - since its input process is the simplest among those handled by the present analysis - to display via (B-18) the dependence of the output rate on the service duration distribution. We do this in the next section.

C3. Another case of special interest in M/G/1/N - the above facility with a finite waiting room. This property can be displayed via the \( \lambda_i \)'s thus

\[
\lambda_n = \begin{cases} 
\lambda, & 0 \leq n < N, \\
0, & n = N.
\end{cases} \quad (C-3)
\]
Here the dependence between \( X_d \) and \( Y \), alluded to in B9, already exists, and thus the methods that can produce \((C-1)\) very simply are inapplicable. However, considerations of the type presented in Appendix D immediately indicate a relationship between the steady-state probabilities of \( X_d \) (let us denote them \( \bar{\pi}_A \)) and those of the associated M/G/1 facility (obtained by letting \( N \to \infty \)). A more complete calculation may be found in [2, III.6], whence one finds
\[
\bar{\pi}_n = \frac{\pi_n}{\sum_{j=0}^{N-1} \pi_j}, \quad 0 < n < N, \tag{C-4}
\]
and further
\[
p(n) = \begin{cases} \frac{\bar{\pi}_n}{(\lambda_1 + \bar{\pi}_0)} & 0 < n < N \\ 1 - \frac{1}{(\lambda_1 + \bar{\pi}_0)} & n = N \end{cases} \tag{C-5}
\]
hence procedures constructed to evaluate \( \pi_i \) need minimal tinkering to be useful here.

C4. We note that the proof of \((B-18)\) as given does not hold for the two last systems, since not all the \( \lambda_i \) differ. But not only is the continuity argument given in Appendix A applicable here, but \((B-18)\) can be directly shown to hold (via the same route, with significant simplifications due to the arrival being now homogeneous Poisson). It is interesting perhaps to note that the markovian state description used in [2] uses attained service rather than the pending requirement.

C5. A somewhat more complex case is the one where the \( \lambda_n \) are a linear function of \( n \). The relation \( \lambda_n = \lambda(N-n)^+ \) gives rise to the so called "machine interference" problem, where there are \( N \) sources, each of
which can either be serviced (one at a time), wait to be serviced, or
request a new service request at the rate $\lambda$. The stochastic processes
that arise in this problem have been rather fully analyzed in [7].

There we find a solution for the limiting distribution of $X_d$, $\tilde{\pi}_i$:

$$\tilde{\pi}_n = \sum_{r=c}^{N-1} (-1)^{r-c} \binom{r}{c} B_r \quad c = N-n-1,$$

$$B_r = C_r \sum_{j=r}^{N-1} \binom{N-1}{j} \frac{1}{C_j} / \sum_{k=0}^{N-1} \binom{N-1}{k} \frac{1}{C_k} ,$$

$$\alpha_i = 1, \quad C_j = \frac{\prod_{i=1}^{j-1} \alpha_i}{1-\alpha_i} , \quad \alpha_i = \mathcal{L}(i\lambda).$$

These $\tilde{\pi}_n$ satisfy equations (B-20).

C6. Again, the relation between $\tilde{\pi}_n$ and the $p$ limiting distribution $p(n)$ is simple:

$$p(n) = \frac{N}{N-n} \frac{\tilde{\pi}_n}{N \alpha x + \tilde{\pi}_0} \quad (C-7)$$

with $p(N)$ given by $1 - \sum_{n=0}^{N-1} p(n)$.

The difficulty in the evaluation of $\tilde{\pi}$ is largely dependent on the
calculation of the $\alpha_i$, and for the extremely versatile family of phase-
type distribution this is a cheap and robust calculation.

In no one of the above cases is a useful expression - i.e., one
open to evaluation by inspection - obtained by substituting the given
solutions into (B-18).
D. DISCUSSION

D1. One of the most commonly used tools in the analysis of the steady state behaviour of birth-and-death systems - i.e., such systems where all intertransition times are exponential - is the arrow diagram; states are represented as points in a 'phase space' with arrows between them indicating permissible transitions and their intensity, or rate. For the univariate version a generic portion of the diagram would be

\[ \begin{array}{c}
\lambda_{n-1} \\
n-1
\end{array} \quad \Rightarrow \quad \begin{array}{c}
\mu_n \\
n
\end{array} \]

and the corresponding local balance equation is \( \lambda_{n-1}p(n-1) = \mu_n p(n) \).

Or \( \mu_n = \lambda_{n-1}p(n+1)/p(n) \). Mark the similarity to (B-18).

D2. Indeed one can reason as follows: suppose a system need be analysed, where not all the participating distributions are exponential. Still, for the purposes of the application in view only a subset of the variables needed for a Markovian description of the process are of interest (such as \( X \) out of \( X \) and \( V \) in Section B). Can still a meaning be attributed to diagrams and balance equations as above? The answer is in the affirmative, provided due care is taken in the estimation and use of the rate functions. What Section B accomplished is a demonstration that the probability of the event "the 'state' is \( n \) and a transition is imminent, given that the 'state' is \( n \)" can be used to manufacture such an intensity function. However, a serious difficulty appears: the \( \mu_n \) of D1 are in certain contexts called "service rates". This shows them for what they are there: a quantity characterizing some service mechanism. When
the service requirements are not exponential. The $r(n)$ do not serve similarly to they not only depend on the service mechanism but on the arrival process as well, or generally - on all the processes that take part in the evolution of the system.

D3. To illustrate this point let us return to (8-18) and view it as a definition of $r(n)$. Further, consider the M/G/1 system - where all $\lambda_n = \lambda$ - which is quite familiar and for which we assume the reader to have some 'feel' as to its likely modes of behaviour. How do the $\pi_n$ and $r(n)$ depend on $n$? When $S \sim \exp(\mu)$ indeed $r(n) = \lambda/\rho = \mu$, a constant. Only when $\pi_n$ is geometrical would this be the case, and only when $S$ is exponential $\pi_n$ is such. Otherwise, all one can say is that $\pi_n$ is a unimodal function of $n$. For not too large $\rho$ the mode is at $n = 0$. For values of $n$ up to, and in the vicinity of, the mode, $r(n)$ may be expected to be rather volatile.

D4. Let us now look at the tail of $\pi_n$, i.e., for $n > \text{mode}(X)$. Here we note that $r(n)$ may be an increasing or decreasing function of $n$, which would correspond to $\pi_n$ being log-concave or log-convex, respectively. We conjecture that $\log \pi_n$ does not have inflection points in the tail, but this was not proved. If it did - which seems unlikely when the input process is homogeneous (and also because of another reason mentioned below) - $r(n)$ would have alternating segments of opposite trends. Log-concavity means $\pi_n$ decrease faster than geometrically, log-convexity implies the decrease is slower than by any constant factor. When will these two modes of behaviour occur? Intuitively - the first happens when the service is more regular than exponential (constant, erlangian, normal with
small $\sigma^2$ etc.) and the other when the service is less regular (DFR, hyper-exponential etc.). Calculations bear this out.

D5. Thus, while the values of $r(n)$ do not bear much evidence towards characterizing the service time (which is confounded with the load), the trend of $r(n)$ characterize the second and higher moments of $S$.

We note however that for the $M/G/1$ system at least, the $r(n)$ are expected to converge to a finite limit. The reason is that for large enough $n$, $\pi_n \sim a \omega^n$, where $a$ is some constant and $\omega$ is the absolute value of the largest eigenvalue of $I - P$, $P$ being the transition matrix given in B-20. Thus $\lim_{n \to \infty} r(n) = \lambda/\omega$. This behaviour is the source of our conjecture on the absence of "special features" in the function $\pi_n$ for values of $n$ which are not small.

D6. The various possible trends of $r(n)$ in conjunction with various service distributions have been observed experimentally (under a somewhat more complicated input process, which can be described within the framework of Section B) [11]. Indeed it was that empirical observation which prompted the present work. There it was observed that when $S$ has a coefficient of variation larger than 1, $r(n)$ decreased sharply with $n$. The converse effect was very weak; no sufficient details were given to explain this difference (it could be due to still higher moments of $S$ or to the shape of the $\lambda_n$). One further remark that is perhaps best placed here is that within the framework of Operational Analysis [5] equations (B-18) - or rather equivalent relationships measured during a finite period in flow-balanced systems - are algebraic identities. This perhaps serves best to exhibit the naturalness of the results of Section B.
D7. This section started by showing that the $r(n)$ can be used in analysis of birth-and-death-like service systems. However, even there we indicated that to do so it had to usurp quantities that characterized the performance of the server. This, we saw, $r(n)$ cannot do: measuring $r(n)$ in one configuration $A$ will not be particularly informative concerning the values $r'(n)$, in $A'$ which has the same server subjected to different load patterns. Since it is the $r(n)$ that appear in the solution of the balance equation (rather then, say, the moments of $S$), prediction is inhibited.

D8. The foregoing give rise to a fine host of speculations concerning the possible use of the $r(n)$ in characterizing possible modes of interactions between service centers and their analysis. The author hopes to report on these in a later work.

ACKNOWLEDGEMENT

The presentation of the results has largely benefitted from discussions with Percy Tzelnick.
REFERENCES


APPENDIX A: EVALUATION OF EQUATION B-9

XA1. The first factor in the RHS of (B-9), \( p(n, v|1, y) \) is the joint density of the probability of hitting the given duration \( Y \) just at \( v \) prior to its termination, and the probability that during a period with duration \( y-v \) the occupancy increased from 1 to \( n \). When \( i=0 \) the increase is from 1 to \( n \). The density component is uniform, with the value \( \frac{1}{y} \). To evaluate the pmf part denote by \( A_k \) a variable \( \exp(\lambda_k) \).

Then, defining \( A_{i, n} = A_i + A_{i+1} + \ldots + A_n \) we may write

\[
y \cdot p(n, v|1, y) = p(A_{i, n-1} < y-v, \ A_{i, n} \geq y-v)
\]

We now make a convenient assumption: all the \( \lambda_k \) are distinct. The case with some \( \lambda_k \) coinciding can be arbitrarily approximated by slightly perturbing these values, thus there is not effective loss in generality, with a substantial simplification. Thus the density of \( A_i, n-1, f_i, n-1(t) \) is given by

\[
a_k(i, n-1) = \prod_{j=1}^{n-1} \lambda_j / \prod_{r=i, r \neq k}^{n-1} (\lambda_r - \lambda_k)
\]

and

\[
y \cdot p(n, v|1, y) = \sum_{k=1}^{n-1} a_k(i, n-1) (e^{-\lambda u} - e^{-\lambda_k u}) (\lambda_k - \lambda_n) \ u = y-v
\]

XA2. The second factor of (B-9), \( P(D=i, Y \leq y) \) is best written as

\[
P(Y \leq y|D=i)P(D=i), \text{ with } Y \text{ depending on } D \text{ only to the extent whether it is zero or not. By considerations similar to those that produced (B-8) it is simple to see that}
\]

\[
d_y P(Y \leq y|D=i) = \begin{cases} \frac{yf(y)}{\alpha} & i > 0 \\ \lambda_0 yf(y)/(1 + \lambda_0 \alpha) & i = 0 \end{cases}
\]
APPENDIX B: DERIVATION OF B-12, THE BALANCE EQUATIONS

XB1. We begin by evaluating $p(n,v; t+\Delta t)$ in terms of the probabilities at time $t$, and will later let $t \to \infty$, with the time derivative of $p(n,v; t)$ vanishing. Writing symbolically

$$p(n,v; t + \Delta t) = ARR + SERV + DEP$$

where for $n=0$ only the last two contribute. Clearly:

$$ARR = \lambda_{n-1} \Delta t \ p(n-1,v; t)(1-\delta_{n,1}) + \lambda_{0} \Delta t f(v) p(0; t) \delta_{n,1},$$

$$SERV = \text{Prob(no arrival)} p(n,v+\Delta t; t),$$

using Taylor's series to first order we get

$$SERV = (1-\lambda_{n} \Delta t)(p(n,v; t) + \Delta t \frac{\partial}{\partial v} p(n,v; t)),$$

and to first order in $\Delta t$

$$SERV = p(n,v; t) + \Delta t \left( \frac{\partial}{\partial v} p(n,v; t) - \lambda_{n} f(n,v; t) \right).$$

$$DEP = \text{Prob (Service terminated during } (t,t+\Delta t); \text{ the subsequent service duration is } v)$$

$$\Delta t = \int_{u=0}^{\Delta t} p(n+1,u; t) du f(v),$$

using the mean value theorem, and anticipating the limit $\Delta t \to 0$, we have

$$DEP = \Delta t p(n+1,0^+) f(v), \ p(n+1,0^+) = \lim_{u \to 0} p(n+1,u).$$

Adding up these contributions, letting $\Delta t \to 0$ and then $t \to \infty$ we obtain for $n > 0$

$$\lim_{t \to \infty} \frac{\partial}{\partial t} p(n,v; t) = 0 = \lambda_{n-1} p(n-1,v)(1-\delta_{n,1}) + \lambda_{0} f(v) p(0) \delta_{1,n} + \frac{\partial}{\partial v} p(n,v) - \lambda_{n} p(n,v) + p(n+1,0^+) f(v).$$

(XB-5)
Collecting the first two terms, writing $p(0,v)$ for $p(0)f(v)$, we get the first part of B-12. The second, for $n=0$, is obtained identically, except that no arrivals contribute.
APPENDIX C: EVALUATION OF (B-15) - \( p(n,0^+) \)

**XC1.** The calculation is simply to substitute (B-10) and (B-11), with (XA-2) in (B-12). We also need the derivative of (B-11)

\[
\frac{\partial}{\partial u} p(n|u) = \begin{cases} 
\sum_{k=1}^{n-1} a_k^{(i,n-1)} \left( \lambda_k e^{-\lambda_k u} - \lambda_n e^{-\lambda_n u} \right) / (\lambda_k - \lambda_n) & \text{if } i < n \\
-\lambda_n e^{-\lambda_n u} & \text{if } i = n
\end{cases}
\]

\[(XC-1)\]

The structure of (XB-5) suggests we first tackle the case \( n = 1 \):

\[
f(v)p(2,0^+) = \lambda_1 p(1,v) - \lambda_0 p(0)f(v) - \frac{\partial}{\partial v} p(1,v).
\]

\[(XC-2)\]

Using (B-10), (B-11) and (XC-1) yields, after \( \frac{\partial}{\partial v} p(1,v) \) is integrated by parts and cancelling \( f(v) \):

\[
p(2,0^+) = \theta(\pi_0 + \pi_1) - \lambda_0 p(0).
\]

\[(XC-3)\]

This expression is further simplified in B8.

**XC2.** For \( n > 1 \) we need evaluate

\[
f(v)p(n+1,0^+) = \lambda_n p(n,v) - \lambda_{n-1} p(n-1,v) - \frac{\partial}{\partial v} p(n,v).
\]

\[(XC-4)\]

Writing, \( \zeta_i = \pi_i + \delta_i, 1 \pi_0 \) the RHS of (XC-4) is

\[
\theta \left( \lambda_n \zeta_n \int_{u=0}^{\infty} f(u+v)e^{-\lambda_n u} du + \zeta_n \int_{u=0}^{\infty} f(u+v)e^{-\lambda_n u} du \right)
\]

\[
+ \sum_{i=1}^{n-1} \int_{u=0}^{\infty} [f(u+v)\{\lambda_n p(n|u)-\lambda_{n-1} p(n-1|u)\}+f'(u+v)p(n|u)]du \}
\]

\[(XC-5)\]
Performing integration by parts on those terms that contain $f'(u+v)$, noting that $p(n|1,0)$ vanishes for $i < n$, transforms the above expression to

$$
\theta \left\{ f_n(v) + \sum_{i=1}^{n-1} \int_{0}^{\infty} f(u+v)[\lambda_n p(n|i,u) - \lambda_{n-1} p(n-1|i,u) + \frac{\partial}{\partial u} p(n|i,u)]du \right\}.
$$

(XC-6)

**XC3.** Taking separately the term in the square brackets of (XC-6) that corresponds to $i = n-1$, we obtain, using (B-11)

$$
\lambda_n a_{n-1} \left[ \frac{e^{-n}u - e^{-n-1}u}{\lambda_{n-1} - \lambda_n} - \lambda_{n-1}u + a\frac{\lambda e^{-n-1} - \lambda_{n-1}e^{-n-1}}{\lambda_{n-1} - \lambda_n} \right].
$$

(XC-7)

Noting from (XC-2) that $a_{n-1} = \lambda_{n-1}$, we see that (XC-7) vanishes. Now take the square brackets in (XC-6) that correspond to some $i, 1 \leq i \leq n-2$:

$$
\sum_{k=1}^{n-2} \left\{ \frac{\lambda_n a_{n-1}}{\lambda_{n-1} - \lambda_k} a_k \left[ \frac{e^{-n}u - e^{-n-1}u}{\lambda_k - \lambda_n} - \lambda_{n-1} a_k \frac{e^{-n-1}u - e^{-n-1}u}{\lambda_k - \lambda_{n-1}} \right] 
\right. 
\left. + \frac{\lambda_n a_{n-1}}{\lambda_{n-1} - \lambda_k} a_k \frac{e^{-n-1}u - \lambda_{n-1}e^{-n-1}}{\lambda_k - \lambda_{n-1}} \right\}.
$$

(XC-8)
Rearranging over a common denominator gives the value of (XC-8) as

\[ \sum_{k=1}^{n-1} a_k^{(i,n-1)} \cdot \lambda_{n-1}^k = \frac{\lambda_{n-1}^{(i,n-1)} - \lambda_{n-1}^{(i,n-2)}}{\lambda_{n-1} - \lambda_k} \text{ for } 1 \leq k \leq n-2. \]

n-1 \sum a_k^{(i,n-1)} \cdot \lambda_{n-1}^k \text{, where the factor } \frac{\lambda_{n-1}}{(\lambda_{n-1} - \lambda_n)} \text{ was reabsorbed in the } a_k^{(i,n-1)} \text{. Using (XA-2) then yields for (XC-8) the value}

\[ \left( \frac{-\lambda_{n-1}^{(i,n-1)} \cdot \lambda_{n-1}^{(i,n-2)}}{\prod_{j=1}^{n} \lambda_j \sum_{k=1}^{n-1} \frac{1}{\prod_{r=1}^{n} (\lambda_r - \lambda_k)}} \right). \]

The sum in (XC-9) vanishes. The simplest way perhaps to see it is to write it with a common denominator, which is \( \prod_{r \neq k} (\lambda_r - \lambda_k) \), and a numerator \( \sum_{k=1}^{n-1} (-1)^{i-k} \sum_{u<v} (\lambda_u - \lambda_v) \). This sum has the value of a determinant of a matrix with row 1 having the value \( (1 \lambda_{x_1} \lambda_{x_2} \cdots \lambda_{x_{n-2}}) \), i.e., a matrix with two (the first ones) identical columns. Thus the square brackets in (XC-6) vanish for all \( i \), and we finally have

\[ f(v)p(n+1,0^+) = \delta \epsilon f(v). \]

Since (XC-4) was evaluated for \( n > 1 \) we obtained (8-15).
APPENDIX D: DERIVATION OF (B-16)

In order to relate \( \pi_n \) and \( p(n) \) we make the following observations:

1. Consider the evolution of the system for a (long) time \( T \). During that time let us suppose there were \( N_a \) arrivals and \( N_d \) departures. Since \( \alpha = E(s) \) is finite, \( (N_a - N_d)/N_a \to 0 \) when \( T \to \infty \).

2. Because arrivals and departures occur singly, the number of arrivals that found \( n \) customers in the system, \( a(n) \), is equal to within \( \pm 1 \) to \( b(n) \), the number of departures that left the system with \( N \) customers. Since \( b(n)/N_d \to \pi_n \), we find that this is also the probability an arrival finds the system with \( n \) customers in it.

3. Let \( T(n) \) be the time during \( T \) in which the system holds precisely \( n \) customers. With probability one (when \( T \to \infty \)) \( a(n) = \lambda_n T(n) \), or \( T(n) = a(n)/\lambda_n \).

4. Since \( p(n) = \lim T(n)/T \), we have
   \[
   p(n) = \lim_{T \to \infty} a(n)/\lambda_n T = \lim_{T \to \infty} N_d / N_a \lambda_n \cdot
   \]

5. With probability one (as \( T \to \infty \)) \( T = T(0) + \alpha N_d \), and since \( p(0) = \lim T(0)/T \) we have that with probability one (as \( T \to \infty \) and when \( p(0) > 0 \))
   \[
   T(0) = \alpha N_d \frac{p(0)}{1-p(0)}, \quad \text{and thus} \quad T = \frac{\alpha N_d}{1-p(0)}.
   \]

6. Substituting this value of \( T \) into the result of 4) and using the closeness of \( N_a \) and \( N_d \) yields
   \[
   p(n) = \lim_{T \to \infty} \pi_n N_a / \lambda_n (\alpha N_d / (1-p(0))) = (1-p(0)) \pi_n / \alpha \lambda_n.
   \]

7. Using the same expression for \( p(0) \) and reinserting in 6) gives the relation (B-16).