R-531-NYC/HUD
MARCH 1971

THE NUMBER OF EMERGENCY UNITS BUSY AT ALARMS WHICH REQUIRE MULTIPLE SERVERS

Jan M. Chaiken

This study was sponsored by the City of New York and by the U. S. Department of Housing and Urban Development. Its contents, however, do not purport to represent the official views or policy of its sponsors.

THE
NEW YORK CITY
RAND
INSTITUTE

545 MADISON AVENUE NEW YORK NEW YORK 10022 (212) 758-2244

Rand

Published by The Rand Corporation

PREFACE

This work is part of an Institute study of the deployment of urban emergency vehicles. We present here the mathematical proof of a formula which can be used to determine how many vehicles to have on duty in particular geographical areas at various times of day. An example of an application of the model, showing the method step by step, is given in a companion report, "Estimating the Number of Fire Engines Needed in New York City Fire Divisions," to appear.

ABSTRACT

Calls for service arrive at an infinite-server queue according to a mixture of Poisson processes. Service for each process occurs in a number of independent stages; stages are identified by the number of emergency units busy serving the call. Assuming arbitrary finite-mean service-time distributions, the distribution of the number of busy units at any time is determined, and the approach to a steady-state distribution is proved.

ACKNOWLEDGMENTS

This work was supported by the New York City Fire Department and the U. S. Department of Housing and Urban Development. I have had helpful discussions about this problem with E. Blum, G. Carter, E. Ignall, R. Larson, P. Pierre, and J. Rolph. I. Greenberg suggested changes in an earlier draft.

CONTENTS

I. INTRODUCTION

We consider a region of a city in which calls for service (alarms) of various types are being generated according to Poisson processes. In response to each alarm, a number of units (such as fire engines, police patrol cars or ambulances) will be dispatched, and additional units may be required as the service progresses. Later, the units may be released from service one at a time or in groups of two or more, until finally the service is complete.

Assuming that the alarm and service rates are not varying with time, we determine the distribution of the number of busy units and investigate the approach to steady-state. We assume that infinitely many units are available, since most emergency services, especially fire departments, will send as many units as are required, even if it is necessary to enlist the assistance of nearby cities to obtain a sufficient number.

We also assume that the service times for each alarm are independent of the number of units already assigned to previous alarms. This assumption restricts our ability to represent travel times realistically. Of course, the observed distribution of travel times can be included in the service-time distribution, but this approach will fail to account for the increase in travel times which generally occurs when many units are busy.

The alarms are separated into types according to the number of units required to serve the alarm and the distributions of the service times for each stage of service. A stage of service consists of a period of time when a fixed number of units are committed to serving the alarm. Thus, a new stage begins whenever an additional unit is dispatched or a unit completes service. By permitting some stages to have zero duration, we allow for the possibility that units may be dispatched or released in groups of two or more.

We assume that the stages of service are independent, and that the number of units engaged increases monotonically to the maximum and then decreases monotonically to zero. Thus, if a maximum of n units is to be assigned to the alarm, there are 2n-1 independent stages of service, which we may number 1,2,..., 2n-1. If $j<n$, the number of units busy in the jth stage is j; if $j>n$, the number of units busy in the jth stage is 2n-j.

To illustrate the stages, we consider one type of fire alarm, which is small enough to be extinguished by the men on two fire engines. We assume that initially three fire engines are dispatched to the scene, so that the first two stages (with one unit busy and two units busy) have zero duration, and stage 3 begins when the alarm arrives. After a period of time, it is determined that two engines are adequate to fight the fire, and the third is released. At this time service enters stage 4, with two units busy. When the fire is extinguished, one of the two remaining engines is released, while the other remains for some follow-up activities, called "overhaul," and this constitutes stage 5.

Since the units perform certain activities together as a group, it is not appropriate to think of the service times of the various units as being independent of each other. However, since the stages constitute distinguishable activities (stage 3 = dispatch, stage 4 = extinguishment, stage 5 = overhaul), it is possible to select the alarm types in such a way that independence of the stages becomes a satisfactory approximation.

The present work constitutes a generalization of Erlang's formulas [2], which were proved valid for the case of an infinite-server queue and general service times by Khintchine [3]. Our results differ from those of Mejzler [4], who considered a system with the same arrival properties as ours, because, in our terminology, he assumed that the service times of the units were mutually independent.

Application

Our results can be used to decide how many emergency units to locate in any specified geographical region. The precise method of application is described elsewhere [1]. In essence, one calculates the probability that more than n units will be busy at once, $n=0,1,2,\ldots$.

Then one assigns enough units to the region so that this probability does not exceed a certain threshold. This method has been used by the New York City Fire Department.

II. EXPONENTIAL SERVICE TIMES

The main results will be derived first for the case where each stage has an exponential holding time. The purpose of this derivation is to describe the essence of the argument without introducing all the complications of the final result. Let us begin with a single alarm type which arrives according to a Poisson process with rate parameter λ and requires, at its peak, n units.

The stages are numbered 1,2,..., 2n-1, and in stage j the number of busy units is $\nu_j = n - |n-j|$. We are assuming that there are positive constants $\mu_1, \ldots, \mu_{2n-1}$ such that the probability of stage j lasting for time t or less is $1 - \exp(-\mu_j t)$.

Under these assumptions, the infinite-server queue can be represented as a Markov process $\{\underline{k}(t): -\infty < t < \infty\}$, where $k_j(t)$ is the number of alarms which are in stage j at time t, j=1,..., 2n-1. Following the method of Morse [5, Chapter 5], we see that the equations of detailed balance for $p(\underline{k})$, the steady-state probability of state $\underline{k}$, are

$$
\begin{aligned}
(\lambda + \Sigma k_j \mu_j) p(k_1, \ldots, k_{2n-1}) &= \lambda\, p(k_1 - 1, k_2, \ldots, k_{2n-1}) \\
&+ (k_1+1)\mu_1\, p(k_1+1, k_2-1, k_3, \ldots, k_{2n-1}) \\
&+ (k_2+1)\mu_2\, p(k_1, k_2+1, k_3-1, k_4, \ldots, k_{2n-1}) \\
&+ \cdots + (k_{2n-1}+1)\mu_{2n-1}\, p(k_1, \ldots, k_{2n-2}, k_{2n-1}+1),
\end{aligned}
$$

where any p with a negative argument is interpreted as zero. The solution to these equations which satisfies the condition that the total probability is 1 is easily seen to be

$$
(1) \qquad p(k_1, \ldots, k_{2n-1}) = e^{-r} \prod_{j=1}^{2n-1} \frac{r_j^{k_j}}{k_j!} ,
$$

where $r_j = \lambda/\mu_j$ and $r = \Sigma r_j$.

The remaining calculations depend only on equation (1) and the fact that the number of units busy in state j is $\nu_j = n - |n-j|$; none of the other assumptions is used in the remainder of the argument. To determine the probability that m units are busy in this process, we observe that when the state is $\underline{k}$, the number of busy units is $\Sigma k_j \nu_j = \underline{k}\cdot\underline{\nu}$. Thus, the probability that m units are busy is

$$P(m) = \sum_{\underline{k}\cdot\underline{\nu}=m} p(\underline{k}).$$

The sum is taken over all non-negative integral values of $k_1,\ldots,k_{2n-1}$ such that $\underline{k}\cdot\underline{\nu} = m$.

Referring back to equation (1), we see that P(m) is the coefficient of s^m in the generating function

$$\Psi(s) = \sum e^{-r} \prod_{j=1}^{2n-1} \frac{(r_j s^{\nu_j})^{k_j}}{k_j!} .$$

The sum is over all non-negative integral values of $k_1,\ldots,k_{2n-1}$. Performing the sum, we have

$$\Psi(s) = \exp\ (-r + \sum_{j=1}^{2n-1} r_j s^{\nu_j}).$$

This can be simplified by setting $\rho(j) = r_j + r_{2n-j}$ for $j=1,\ldots, n-1$, and $\rho(n) = r_n$. For convenience, we also set $\rho(j) = 0$ for $j=0$ or $j>n$. Then we have

$$(2) \qquad \Psi(s) = \exp\left[-r + \sum_j \rho(j)s^j\right]$$

$$= e^{-r}\left[1 + \Sigma\rho(j)s^j + \frac{1}{2!}(\Sigma\rho(j)s^j)^2 + \cdots\right].$$

The sum is over all non-negative integral values of j, but is actually a finite sum.

From equation (2) we see that the coefficient of s^m in the Taylor series expansion for $\Psi(s)$ is given by

$$P(m) = e^{-r} \text{ if } m=0$$

$$P(m) = e^{-r}\left\{\rho(m) + \frac{1}{2!}\rho^{*2}(m) + \frac{1}{3!}\rho^{*3}(m) + \cdots + \frac{1}{m!}\rho^{*m}(m)\right\},$$

for m>0. Here the asterisk stands for convolution:

$$\rho^{*2}(m) = \rho*\rho(m) = \sum_{j=0}^{m} \rho(m-j)\rho(j),$$

and

$$\rho^{*k}(m) = \rho*\rho^{*k-1}(m) \text{ for } k>2.$$

The sum for P(m) is finite because $\rho^{*k}(m) = 0$ when k>m.

This result is easily generalized to the case of several independent types of alarms. We state the result formally as a proposition which will be generalized in the theorem in the next section.

PROPOSITION 1

Suppose the input to an infinite server queue consists of ℓ independent Poisson processes having rates $\lambda_1, \ldots, \lambda_\ell$. Suppose further that the process with rate λ_i requires $2n_i-1$ independent stages of service, such that the jth stage has an exponential holding time with finite, non-zero mean $T_i(j)$ and requires $n_i - |n_i - j|$ servers, $j=1, \ldots, 2n_i-1$. Then the steady-state probability that m servers are busy is

$$P(m) = e^{-r} \text{ for } m=0$$

and

$$(3) \qquad P(m) = e^{-r} \left[\rho(m) + \frac{1}{2!}\rho^{*2}(m) + \cdots + \frac{1}{m!}\rho^{*m}(m)\right]$$

for m>0. Here $\rho(j) = \sum_{i=1}^{\ell} \rho_i(j)$, with

$$(4) \qquad \rho_i(j) = \begin{cases} \lambda_i\,(T_i(j) + T_i(2n_i-j)) & \text{if } j=1, \ldots, n_i-1 \\ \lambda_i T_i(n_i) & \text{if } j=n_i \\ 0 & \text{otherwise} \end{cases}$$

and $r = \sum_j \rho(j)$.

Proof: If $P_i(k)$ denotes the probability that k units are busy at alarms of type i, and $\Psi_i(s) = \sum_k P_i(k)s^k$, then the probability that, in total, m units are busy is the coefficient of s^m in the generating function

$$\Psi(s) = \prod_{i=1}^{\ell} \Psi_i(s).$$

Using (2) we have

$$\Psi(s) = \exp\left[\sum_{i=1}^{\ell} \left(-r_i + \sum_j \rho_i(j)s^j\right)\right],$$

where $\rho_i(j)$ is given by (4), and $r_i = \sum_j \rho_i(j)$. Performing the sum on i, we obtain

$$\Psi(s) = \exp\left(-r + \sum_j \rho(j)s^j\right).$$

The same argument used earlier shows that the coefficient of s^m in the Taylor series for $\Psi(s)$ is given by (3). This completes the proof of the proposition.

To apply this result using data from actual alarms, it is sometimes convenient to determine the average service time separately for each stage of each type of alarm. However, a simpler method can be used if we observe that $\rho(j) = \lambda\tau(j)$, where $\tau(j)$ is the average, over all alarms of all types, of the time that exactly j units spend working together. This can be seen by rewriting the equation for $\rho(j)$ as

$$\rho(j) = \lambda \sum_{i=1}^{\ell} \frac{\lambda_i}{\lambda} \tau_{ij},$$

where τ_{ij} is the mean total time spent by exactly j units at an alarm of type i.

III. GENERAL SERVICE-TIME DISTRIBUTIONS

Proposition 1 remains valid if we omit the hypothesis that the service-time distributions are exponential and assume instead that each service-time distribution has a finite mean. For certain special distributions, the generalization can be derived using a variant of the proof in Section II:

- For Erlang distributions, which may be considered as staged exponential distributions, and for zero-duration distributions, we can increase or decrease the number of stages and modify the number ν_j of units busy in stage j. The result remains the same.
- For distributions which are probabilistic mixtures of exponential, Erlang, and zero-duration distributions, we can increase the number of alarm types, so that each alarm type receives service in stages having one of these three types of distributions.

For more general distributions, we shall demonstrate that the probability of finding m units busy at time t approaches P(m), as given in equation (3), independent of the initial state of the system. For this it suffices to prove, for each alarm type, that the probability $p(\underline{k};t)$ of finding k_j alarms in stage j at time t approaches the value given in equation (1).

PROPOSITION 2

Suppose that an infinite-server queue receives alarms according to a Poisson process with rate λ. Suppose further that each alarm is served in N stages, with

$$H_j(t) = \Pr\left\{\text{duration of stage } j \text{ is } \leq t\right\}$$

satisfying

$$T_j = \int_0^\infty t\,dH_j(t) < \infty, \; j=1,\ldots,N.$$

Given any distribution of the number of alarms in each stage at time $t = 0$, let $p(k_1,\ldots, k_N;t)$ represent the probability that at time $t \geq 0$ the number of alarms in stage j is k_j, $j=1,\ldots, N$. Then

$$\text{(5)} \qquad \lim_{t\to\infty} p(k_1,\ldots, k_N;t) = e^{-r} \prod_{j=1}^{N} \frac{r_j^{k_j}}{k_j!} ,$$

where $r_j = \lambda T_j$ and $r = \sum_{j=1}^{N} r_j$.

Remarks: This result is somewhat surprising, since only the means of the service-time distributions enter into the steady-state solution, and, moreover, the same distribution of the states would be obtained if each stage arrived according to a Poisson process with rate λ, independent of the arrival times of the other stages. The Proposition can be proved directly from Khintchine's result [3], but the method given below contains some interesting intermediate results.

Proof: We assume first that no alarms are being served at $t = 0$. Suppose further that n alarms have arrived in the interval $[0,t]$. Denote the arrival times by $t_1,\ldots, t_n$, where each t_i is uniformly distributed over $[0,t]$, independent of t_k for $k\neq i$. The probability $P_i(j,t)$ that the alarm which arrived at t_i is in stage j of service at time t is the probability that the sum of the service times of the first j-1 stages is less than $t-t_i$ and service is not complete on stage j at t. Thus

$$\text{(6)} \qquad P_i(j,t) = \int_{s=0}^{t-t_i} d\left(h_1 * \cdots * h_{j-1}\right)(s)\ (1 - H_j(t-s-t_i)),$$

where h_m denotes the measure whose cumulative distribution is H_m, and the asterisk denotes convolution of the measures. The probability that an arbitrary alarm is in stage j of service is then

(7) $$p(j,t) = \frac{1}{t}\int_0^t P_i(j,t)dt_i.$$

Hence, given that n alarms arrived independently, uniformly distributed in $[0,t]$, the probability of finding k_j in stage j at time t, $j=1,\ldots, N$, is

$$p_n(k_1,\ldots, k_N;t) = \frac{n!}{k_1!\cdots k_N!(n-\Sigma k_i)!}\left(\prod_{j=1}^{N} p(j,t)^{k_j}\right)(1-\Sigma p(i,t))^{n-\Sigma k_i}$$

(for $\Sigma k_j \leq n$).

Defining $p_n(k_1,\ldots, k_N;t) = 0$ if $\Sigma k_j > n$, and using the assumption of Poisson arrivals, we can allow an arbitrary value for n, and we find that the unconditional probability of finding k_j alarms in stage j at time t, $j=1,\ldots, N$, is

$$p(k_1,\ldots, k_N;t) = \sum_{n=0}^{\infty} \frac{(\lambda t)^n}{n!} e^{-\lambda t} p_n(k_1,\ldots, k_N;t)$$

$$= \sum_{n=\Sigma k_i}^{\infty} (\lambda t)^n e^{-\lambda t} \left(\prod_{j=1}^{N} \frac{p(j,t)^{k_j}}{k_j!}\right) \frac{(1-\Sigma p(i,t))^{n-\Sigma k_i}}{(n-\Sigma k_i)!},$$

or

(8) $$p(k_1,\ldots, k_N;t) = \prod_{j=1}^{N} \left[\frac{(\lambda t p(j,t))^{k_j}}{k_j!} e^{-\lambda t p(j,t)}\right].$$

Thus we see that the number of alarms in the various stages are independently Poisson distributed, with mean $\lambda t p(j,t)$ in the jth stage. This mean depends on the service-time distributions in all the stages from the first to the jth. This result is of interest in itself.

From equation (8) we see that the proposition will be proved for the case that the queue is empty at $t = 0$, if we can demonstrate that

$$\lim_{t\to\infty} tp(j,t) = T_j .$$

But from equations (6) and (7) we have

$$tp(j,t) = \int_{t'=0}^{t} dt' \int_{s=0}^{t-t'} d\alpha_j(s)\ (1 - H_j(t-s-t'))$$

$$= \int_{u=0}^{t} du \int_{s=0}^{u} d\alpha_j(s)\ (1 - H_j(u-s))$$

where α_j is the measure $h_1 * \cdots * h_j$. The limit as $t \to \infty$ is simply the Fourier transform of the convolution on the right, evaluated at zero. However, since

$$\int_{o}^{\infty} (1 - H_j(t))\ dt = T_j, \text{ and } \int_{o}^{\infty} d\alpha_j = 1,$$

and the Fourier transform of a convolution is the product of the Fourier transforms, we see that

$$\lim_{t\to\infty} tp(j,t) = T_j .$$

Thus the proposition is proved for the case that the system is empty at $t = 0$. The generalization to the case of an arbitrary initial distribution of the alarms in each stage can be proved following the method of Mejzler [4, Section 5].

From Proposition 2 and equation (7), it is easy to prove the following generalization of Proposition 1, and we omit the proof.

THEOREM

Suppose the input to an infinite-server queue consists of independent Poisson processes having rates $\lambda_1, \lambda_2, \lambda_3, \ldots$. Suppose further that the process with rate λ_i requires $2n_i-1$ independent stages of service, with the jth stage having a holding time distribution H_{ij} with a finite mean $T_i(j)$ and requiring $n_i - |n_i - j|$ servers. We assume that $r = \sum_{i,j} \rho_i(j) < \infty$, where $\rho_i(j)$ is given by (4).
Given any distribution of the number of alarms in each stage of each alarm type at $t = 0$, let $P(m,t)$ be the probability that m servers are busy at time $t \geq 0$. Then

$$\lim_{t\to\infty} P(m,t) = P(m),$$

where $P(m)$ is as given in Proposition 1.

Furthermore, if no servers are busy at $t = 0$, the probability that m servers are busy at time t is

$$P(m,t) = e^{-r(t)} \text{ for } m = 0$$

and $$P(m,t) = e^{-r(t)} \left(\rho(m,t) + \frac{1}{2!}\rho^{*2}(m,t) + \cdots + \frac{1}{m!}\rho^{*m}(m,t)\right)$$

for $m > 0$. Here $\rho(j,t) = \sum_i \rho_i(j,t)$ and $r(t) = \sum_j \rho(j,t)$, with

$$\rho_i(j,t) = \begin{cases} \lambda_i \tau_i(j,t) + \lambda_i \tau_i(2n_i-j,t) & j=1,\ldots, n_i-1 \\ \lambda_i \tau_i(n_i,t) & j=n_i \\ 0 & \text{otherwise} \end{cases},$$

$$\tau_i(j,t) = \int_{s=o}^{t} \int_{u=o}^{t-s} d\left(h_{ij} * \cdots * h_{ij-1}\right) (s)\ (1 - H_{ij}(u))\ du,$$

and h_{ij} is the measure such that

$$H_{ij}(t) = h_{ij}([0,t]).$$

REFERENCES

1. Chaiken, J.M., "Estimating the Number of Fire Engines Needed in NYC Fire Divisions," R-508, New York City-Rand Institute, to appear.

2. Erlang, A.K., Post Office Electrical Engineer's Journal, 10, 189-197 (1917).

3. Khintchine, A. Ya., Mathematical Methods in the Theory of Queueing, Charles Griffin, London (1960).

4. Mejzler, D., J. Appl. Prob. 5, 143-157 (1968).

5. Morse, P.M., Queues, Inventories, and Maintenance, John Wiley, New York (1967).