QUEUES WITH ORDERED SERVERS THAT WORK AT DIFFERENT RATES

Robert B. Cooper

New Mexico Institute of Mining and Technology, New Mexico 87801, USA

(Received : March 1976, Revised : June 1976)

ABSTRACT

We consider a queuing system in equilibrium in which customers arrive according to a Poisson process and request service from a group of parallel heterogeneous exponential servers that are numbered 1, 2,.., s. An arriving customer is served by the lowest-numbered idle server. The parameter of the exponential distribution that characterizes the jth server depends on the index j; that is, each server works at its own characteristic rate. A blocked customer (one who finds all s servers busy) may defect from the system or wait in accordance with any scheme such that the states of the system comprise a birth-and-death process. For this model we calculate (*i*) the probability that an arriving customer will find all s servers busy and $i = 0, 1, \ldots$. other customers waiting in the queue, and (*ii*) the load carrried by (utilization of) each server in the ordered group. These quantities sometimes permit straight-forward calculation of other important quantities, such as the waiting time distribution function when all blocked customers wait until served and service is in order of arrival. The method of solution is of interest in itself; it consists of recognizing that the complicated set of multidimensional birth-and-death equations that describe this model are essentially the same as those that describe a different model, which in turn can be analyzed by a method that does not require solution of these equations.

1. Introduction

We consider an equilibrium queuing system composed of s heterogeneous exponential servers, numbered 1, 2, ..., s, in which each arriving customer is served by the lowest-numbered idle server. The queuing system is represented as a multidimensional birth-and-death process as follows: Let λ_k and μ_k ($k = 0, 1, \ldots$) be the birth and death rates when the total number of customers present (waiting or in service) is k. We require only that $\lambda_k = \lambda$ when $k < s$ (customers arrive in a Poisson stream when there is at least one idle server) and $\mu_k = \mu(1)x_1 + \ldots + \mu(s)x_s$ when $k \leq s$, where $[\mu(j)]^{-1}$ is the mean service time for the jth server and x_j is the realization of a random variable X_j that assumes the value 0 when the jth server is idle

and 1 when it is busy. The birth rates $\lambda_k (k \geqq s)$ and death rates $\mu_k (k > s)$ can be arbitrarily chosen in the usual fashion to correspond to different arrival processes and queue disciplines (for example, if the waiting room can accommodate at most n customers, then $\lambda_k = 0$ when $k \geqq s + n$). For this model we calculate (*i*) the probability that an arriving customer will find all s servers busy and $i = 0, 1, \ldots$ other customers waiting in the queue, and (*ii*) the probability $P\{X_j = 1\}$ that the jth ordered server is busy (that is, the utilization of or load carried by the jth server). These quantities, which are easily calculated, permit straightforward calculation of other important quantities in certain models, such as the waiting time distribution function for order-of-arrival service when all blocked customers wait until served.

The method of solution is of interest in itself; it consists of recognizing that the complicated set of multidimensional birth-and-death equations that describe this model are essentially the same as those that describe a different model, which in turn can be analyzed by a method that does not require the solution of these equations.

In the standard birth-and-death queuing model, in which the servers are homogeneous (statistically identical), the system in equilibrium can be described in the usual manner by the well known one-dimensional birth-and-death equations that relate the rates at which the system moves from state to state. In this case the *state* of the system is taken simply as the number of customers present without regard to which servers they occupy. Thus the order of search for an idle server is irrelevant unless one is interested in the behaviour or states of a particular server instead of only the behaviour of the system as a whole. On the other hand, if the servers work at different rates, then a description of the (birth-and-death) process requires specification of not only the total number of customers present, but also the identities of the particular servers they occupy. Thus, the way in which a customer chooses which server to occupy when more than one is available must be specified when the servers are heterogeneous. And the birth-and-death equations that describe the case of heterogeneous servers are now multidimensional instead of one-dimensional. These multidimensional equations are no more difficult than their one-dimensional counterparts to derive, but they are much more difficult to solve since, unlike the one-dimensional equations, they do not, in general, yield a simple closed-form solution.

Most previous authors who have considered queues with heterogeneous servers have assumed either random or ordered selection of idle servers, and all have approached the problem through solution of the detailed multidimensional equilibrium birth-and-death equations. For the case where an arrival who finds at least two idle servers chooses his server at random from

all those idle, Gumbel [4] has shown that, as sometimes occurs in such problems, the multidimensional birth-and-death equations for this case do admit a simple closed-form solution whose correctness is easily verified by substitution.

On the other hand, if service is provided by heterogeneous servers that are selected in a prescribed order, the problem is more difficult in the sense that the multidimensional birth-and-death equations for this case do not admit a simple closed-form solution. Singh has solved the equilibrium birth-and-death equations by brute force for the special case of $s = 2$ servers (Singh [9], see also pp. 220-223 of [1]) and $s = 3$ servers (Singh [10]). In Ref. [9] Singh was concerned primarily with the question of finding the optimal allocation of service rates between the two servers. (see Ref. [9] also for additional references.) This question has been considered by Tahara and Nishida [12] for the corresponding model with an arbitrary number of servers but with no waiting positions. It will follow as an easy consequence of our analysis that the conclusions of Tahara and Nishida remain valid for the more general model considered here. The interesting model in which each customer minimizes his own expected sojourn time by (possibly) refusing to accept service from an idle server if he can reduce his expected remaining sojourn time by waiting for a faster server to become idle was considered by Godini [3], who set up, but did not solve, the multidimensional birth-and-death equations.

The problem of homogeneous servers that are selected in a prescribed order is of importance in teletraffic theory, particularly in studies of alternate routing of telephone calls and determination of the load carried by each trunk (server) of an ordered group. These efforts, which date from the 1920's are described in Syski [11] under the headings of alternate routing, gradings, hunting, overflow, and limited availability. Recently, Wallström [14] has generalized a classic teletraffic model to include heterogeneous servers. In Wallström's model, customers arrive in a Poisson stream and seek service first from any server in a primary group. If an arrival finds all servers in the primary group busy, he overflows to a secondary group. If he also finds all servers in the overflow group busy, he is cleared from the system without receiving service. The service times provided by servers in the primary group are assumed to be exponentially distributed with mean α^{-1}, while the service times provided by the secondary group are exponentially distributed with (different) mean β^{-1}. Wallström solves the (two-dimensional) equilibrium probability state equations for the joint distribution of the number of customers present simultaneously on each of the two server groups. Also, he obtains expressions for the mean and variance of the marginal distribution of the number of customers on an infinity-server secondary group, results that are of interest in teletraffic applications.

Wallström's model differs from ours in that (*a*) it allows only two different service rates and (*b*) it does not include provision of waiting positions for blocked customers. Kühn [6] discusses numerical methods for analysis of a model that differs from Wallström's in that Kühn's model allows for provision of a finite number of waiting positions. Both of these models are included as special cases of the model presented here. Recently, Forys and Messerli [2] studied a similar model, again with no provision of waiting positions, to describe a telephone trunk group containing faulty trunks (which are characterized by short holding times).

The strategy of the present paper is to attack the problem of ordered heterogeneous servers through application of some ideas that teletraffic theorists have used in studies of ordered homogeneous servers. This approach permits solution of the problem of ordered heterogeneous servers without necessitating the solution of the detailed multidimensional birth-and-death equations.

2. Statement of Results

The model under consideration is the following: Customers request service from a group of servers that are numbered 1, 2, ..., s. An arriving customer is served by the lowest-numbered idle server. No server can be idle if a customer is waiting. The probability that exactly one customer will arrive in any interval of length h during which k is the number of customers present is $\lambda_k h + o(h)$ as $h \to 0$, $k = 0, 1, \ldots$, where $\lambda_k = \lambda$ when $k < s$; and the probability of more than one arrival is $o(h)$ as $h \to 0$, independently of all other considerations. Similarly, the probability that exactly one customer will depart from the system (through service completion or defection from the queue) in any interval of length h during which k customers are present is $\mu_k h + o(h)$ as $h \to 0$, $k = 0, 1, \ldots$, where $\mu_k = \mu(1)x_1 + \ldots + \mu(s)x_s$ when $k \leqslant s$, and where x_j ($j = 1, 2, \ldots, s$) is the realization of a random variable X_j defined by

$$X_j = \begin{cases} 0 \text{ when the } j\text{th ordered server is idle,} \\ 1 \text{ when the } j\text{th ordered server is busy;} \end{cases} \tag{1}$$

also, the probability of more than one departure is $o(h)$ as $h \to 0$. (Note that the death rate $\mu_k (k \leqslant s)$ depends not only on the number k of busy servers, but also on the identities of the busy servers).

We define $\gamma_j(z)$ by the recurrence

$$\gamma_{j+1}(z) = \frac{\gamma_j[z + \mu(j)]}{1 - \gamma_j(z) + \gamma_j[z + \mu(j)]}, \qquad [j = 1, 2, \ldots, s-1; \quad \gamma_1(z) = \lambda/(\lambda + z)] \tag{2}$$

and let

$$B_j = \gamma_1[\mu(1)]\ \gamma_2[\mu(2)] \cdots \gamma_j[\mu(j)]. \qquad (j = 1, 2, \ldots, s). \tag{3}$$

Finally, we define

$$A_i = \begin{cases} 1, & (i = 0) \\ \dfrac{\lambda_s \lambda_{s+1} \cdots \lambda_{s+i-1}}{\mu_{s+1}\mu_{s+2} \cdots \mu_{s+i}}, & (i = 1, 2, \ldots) \end{cases} \tag{4}$$

and

$$A = \sum_{i=1}^{\infty} A_i. \tag{5}$$

Then we have the following results :

(*i*) If P_k is the equilibrium probability that there are k customers in the system (in service or waiting for service) at an arbitrary instant, then $P_k = 0$ $(k = 0, 1, \ldots)$ when $A = \infty$; if $A < \infty$, then

$$P_{s+i} = \frac{B_s}{1 + AB_s} A_i. \qquad (i = 0, 1, \ldots) \tag{6}$$

(*ii*) If p_j is the load carried by (or the utilization of) the jth ordered server (that is, $p_j = P\{X_j = 1\}$), then

$$p_j = \frac{\lambda}{\mu(j)}(B_{j-1} - B_j)\left(1 - \sum_{i=1}^{\infty} P_{s+i}\right) + \sum_{i=1}^{\infty} P_{s+i}. \tag{7}$$

$$(j = 1, 2, \ldots, s;\ B_0 = 1)$$

3. Proof of Results

We begin with the temporary assumption that blocked customers are cleared; that is, there are no waiting positions, so every arriving customer who finds all s servers busy leaves the system immediately. Let

$$\tilde{P}(x_1, \ldots, x_s) = P\{X_1 = x_1, \ldots, X_s = x_s\} \tag{8}$$

be the equilibrium joint probability function of the random variables X_1, $\ldots$, X_s, where X_j $(j = 1, 2, \ldots, s)$ takes the value 0 when the jth server is idle and the value 1 when it is busy. Then, as is well known, these probab lities satisfy the following system of *conservation-of-flow* equations for all $x_j = 0$ or 1 $(j = 1, 2, \ldots, s)$ except $x_1 = \ldots = x_s = 1$:

$$[\lambda + \mu(1)x_1 + \ldots + \mu(s)x_s]\,\tilde{P}(x_1, \ldots, x_s)$$

$$= \lambda \sum_{j=1}^{s} \delta(j, x_1 + \ldots + x_j)\,\tilde{P}(x_1, \ldots, x_{j-1}, x_j - 1, x_{j+1}, \ldots, x_s)$$

$$+ \sum_{j=1}^{s} \mu(j)\,\delta(0, x_j)\,\tilde{P}(x_1, \ldots, x_{j-1}, x_j + 1, x_{j+1}, \ldots, x_s). \tag{9}$$

where, by definition, $\delta(x, y) = 0$ when $x \neq y$ and 1 when $x = y$. Note that the set (9) comprises $2^s - 1$ independent equations for the 2^s unknown probabilities $\{\tilde{P}(x_1, \ldots, x_s)\}$. Another equation, which is valid when $x_1 = \ldots = x_s = 1$, can be obtained from (9) by omitting the term in λ on the left-hand side; this equation can also be obtained as the sum of all the equations in the set (9), and is therefore redundant. The final independent equation required to determine uniquely the distribution $\{\tilde{P}(x_1, \ldots, x_s)\}$ is the normalization equation

$$\Sigma\, \tilde{P}(x_1, \ldots, x_s) = 1, \tag{10}$$

where the summation is carried out over all 2^s probabilities.

We shall now determine, without solving the set (9), the probabilities $\{B_j\}$ defined by

$$B_j = \Sigma\, \tilde{P}\,(1, 1, \ldots, 1, x_{j+1}, x_{j+2}, \ldots, x_s), \qquad (j = 1, 2, \ldots, s) \tag{11}$$

where the summation is extended over all 2^{s-j} probabilities for which $x_1 = \ldots = x_j = 1$. B_j is thus the probability that an arriving customer finds all the first j ordered servers busy.

Observe that the (conditional) probability that a customer who finds the first $j-1$ ordered servers busy also finds the jth ordered server busy is B_j/B_{j-1} ($j=1, 2, \ldots, s$; $B_0=1$). If we let $\gamma_j(z)$ be the Laplace-Stieltjes transform of the distribution function of elapsed time between two successive instants at which an arriving customer *overflows* the first $j-1$ ordered servers and thus requests service from the jth ordered server, then $\gamma_j[\mu(j)]$ is this same probability;

that is,

$$\gamma_j[\mu(j)] = \frac{B_j}{B_{j-1}}. \qquad (j = 1, 2, \ldots, s) \tag{12}$$

Thus, B_j ($j=1, 2, \ldots, s$) is given by (3). Further, it follows from the work of Palm [7] on overflows in systems with no waiting positions (see also pp. 88–91 of [5], pp. 36-37 of [8], and pp. 262-263 of [11]) that the transforms $\gamma_j(z)$ are determined by the recurrence (2).

Next we calculate $\tilde{P}_j = P\{X_j = 1\}$, the load carried by the jth ordered server, which is defined in terms of the probabilities (8) as

$$\tilde{p}_j = \Sigma\, \tilde{p}\,(x_1, \ldots, x_{j-1}, 1, x_{j+1}, \ldots, x_s), \tag{13}$$

where the sum is taken over all values of the arguments

$$x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_s.$$

As was done with the probabilities $\{B_j\}$, we will calculate the distribution $\{\tilde{p}_j\}$ without solving the equations (9).

Observe that, under the assumption that blocked customers are cleared, λB_{j-1} is the rate at which customers overflow the first $j-1$ servers and request service from the jth. Hence $\lambda B_{j-1}-\lambda B_j$ is the rate at which the jth ordered server is seized by arriving customers; and $(\lambda B_{j-1}-\lambda B_j)/\mu(j)$ is therefore the equilibrium proportion of time that the jth server is busy. We conclude that

$$\tilde{p}_j = \frac{\lambda}{\mu(j)}\,(B_{j-1}-B_j). \tag{14}$$

We have now proved the results (*i*) and (*ii*) for the special case in which blocked customers are cleared. Our method did not require solution of the equilibrium probability state equations (9), but instead was based on an argument whose validity depends on the assumption that blocked customers are cleared. We will now drop that assumption and show that, surprisingly, essentially the same results hold, even though the argument made in the derivation of these results no longer applies.

First, let us define

$$P(x_1,\ldots, x_s; i) = P\{X_1{=}x_1,\ldots,X_s{=}x_s; Q{=}i\}, \tag{15}$$

where Q is the number of customers waiting in the queue at an arbitrary instant. Then the probabilities defined in (15) satisfy, for $i=0$, the same set of equations (9) satisfied by the probabilities (8) for the case of blocked customers cleared. From this important observation we can conclude that the probabilities $\{\tilde{P}\}$ and $\{P\}$ are proportional (by c) when $i=0$:

$$P(x_1,\ldots,x_s; 0) = c\tilde{P}\,(x_1,\ldots,x_s). \tag{16}$$

For notational simplicity, let us write

$$P_{s+i} = P(1, 1,\ldots, 1; i). \qquad (i = 0, 1,\ldots) \tag{17}$$

Then the probabilities P_{s+i} $(i=0, 1,\ldots)$ satisfy the recurrence

$$\lambda_{s+i} P_{s+i} = \mu_{s+i+1} P_{s+i+1}, \qquad (i = 0, 1,\ldots) \tag{18}$$

from which we can conclude that

$$P_{s+i} = \frac{\lambda_s\,\lambda_{s+1}\cdots\lambda_{s+i-1}}{\mu_{s+1}\,\mu_{s+2}\cdots\mu_{s+i}}\,P_s. \qquad (i = 1, 2,\ldots) \tag{19}$$

If we define A_i $(i=0,1,\ldots)$ according to equation (4), then equation (19) can be written

$$P_{s+i} = A_i P_s \qquad (i = 0, 1,\ldots) \tag{20}$$

From equations (16) and (20) we can write

$$P_{s+i} = cA_i \, \tilde{P}(1, 1,\ldots,1), \qquad (i = 0, 1,\ldots) \tag{21}$$

which, by virtue of (11), becomes

$$P_{s+i} = cA_i \, B_s. \qquad (i = 0, 1,\ldots) \tag{22}$$

It remains to determine the constant c from the normalization equation

$$\Sigma \, P(x_1,\ldots,x_s; 0) + \sum_{i=1}^{\infty} P_{s+i} = 1, \tag{23}$$

where the summation on the left hand side is extended aver all values of the arguments $x_1,\ldots, x_s$.

Substitution of equation (16) and (22) into (23) gives

$$c \, \Sigma \, \tilde{P}(x_1,\ldots, x_s) + cAb_s = 1. \tag{24}$$

where A is defined by (5). Equations (24) and (10) give

$$c = (1+AB)^{-1}; \tag{25}$$

and equations (22) and (25) yield (6) which completes the proof of result (*i*).

It remains only to calculate $p_j = P\{X_j = 1\}$, the load carried by the jth ordered server. Let N be the total number of customers in the system (in service or waiting in the queue). Then, from the law of total probability, we can write

$$p_j = P\{X_j{=}1 \mid N \leqslant s\} \, P\{N \leqslant s\} + P\{X_j{=}1 \mid N > s\} \, P\{N > s\}. \qquad (j = 1, 2, \ldots, s) \tag{26}$$

Clearly, $$P\{X_j = 1 \mid N > s\} = 1.$$

Vaulot [13] has given a clever intuitive argument that can be adapted to the present case to show that

$$P\{X_j = 1 \mid N \leqslant s\} = \tilde{p}_j. \qquad (j = 1, 2, \ldots, s) \tag{27}$$

that is, the load carried by the jth ordered server during those time intervals when no customers are waiting equals the total load carried by the jth ordered server in the corresponding system in which customers are not allowed to wait. The intuitive argument in support of (27) is simply that for a birth-and-death process the stochastic behaviour of the system during the time intervals when $N \leqslant s$ is unaffected by the stochastic behaviour of the system when $N > s$ Equations (26) and (27), together with (14), yield (7); thus result (*ii*) will be established if we can provide a rigorous proof of (27).

To this end, we apply the definition of conditional probability to write

$$P\{X_j = 1 \mid N \leqslant s\} = \frac{P\{X_j = 1,\ N \leqslant s\}}{P\{N \leqslant s\}} \ . \tag{28}$$

Now

$$P\{X_j = 1,\ N \leqslant s\} = \Sigma\ P(x_1, \ldots, x_{j-1},\ 1,\ x_{j+1}, \ldots, x_s;\ 0), \tag{29}$$

and

$$P\{N \leqslant s\} = \Sigma\ P(x_1, \ldots, x_s;\ 0), \tag{30}$$

where the summations in equations (29) and (30) extend over all values of the arguments $\{x_i\}$. Substitution of (16) into (29) gives, by virtue of (13),

$$P\{X_j = 1,\ N \leqslant s\} = c\tilde{p}_j. \tag{31}$$

Similarly, substitution of (16) into (30) gives, by virtue of (10),

$$P\{N \leqslant s\} = c. \tag{32}$$

Substitution of (31) and (32) into (28) yields (27). The proof is complete.

ACKNOWLEDGEMENT

Mr. Borge Tilt provided a careful reading of the manuscript and made many helpful comments.

REFERENCES

[1] BHAT, U.N. (1972), *Elements of Applied Stochastic Processes*, Wiley, New York.

[2] FORYS, L.J., and MESSERLI, E.J. (1975), Analysis of trunk groups containing short-holding-time trunks, *The Bell System Tech. J.*, **54**(6), 1127-1153.

[3] GODINI, G. (1965), A problem in the theory of queues with heterogeneous servicing, *Studii Cercetari Mathematice*, **17**(5), 765-775 (in Rumanian).

[4] GUMBEL, H. (1960), Waiting lines with heterogeneous servers, *Operations Research*, **8**(4), 504-511.

[5] KHINTCHINE, A.Y. (1969), *Mathematical Methods in the Theory of Queuing*, 2nd ed., Hafner, New York.

[6] KÜHN, P. (1973), The impact of queuing theory on the optimization of communications and computer systems, *Proceedings of the 20th International Meeting of the Institute of Management Sciences*, Tel Aviv, June 24-29, 559-568.

[7] PALM C. (1943), Intensitätsschwankungen im fernsprechverkehr, *Ericsson Technics*, **44**, 1-189.

[8] RIORDAN, J. (1962), *Stochastic Service Systems*, Wiley, New York.

[9] SINGH, V.P. (1970), Two-server Markovian queues with balking : heterogeneous vs. homogeneous servers, *Operations Research*, **18**(1), 145-159.

[10] SINGH, V.P. (1971), Markovian queues with three heterogeneous servers, *AIIE Transactions*, 3(1), 45-48.

[11] SYSKI, R. (1960), *Introduction to Congestion Theory in Telephone Systems*, Oliver and Boyd, Edinburgh.

[12] TAHARA, A. and NISHIDA, T. (1975), Optimal allocation of service rates for multi-server Markovian queue, *J. Operations Research Soc. of Japan*, **18** (1&2), 90-96.

[13] VAULOT, E. (1925), Application du calcul des probabilités a l'exploitation telephonique, *Annales des Postes, Télégraphes et Téléphones*, **14**(2), 136-156.

[14] WALLSTRÖM, B. (1973), Loss calculations in certain overflow systems where the holding times in successive groups have different means, *Proceedings of the Seventh International Teletraffic Congress*, Stockholm, June 13-20, paper 417.