J. Appl. Prob. **20**, 144–152 (1983)
Printed in Israel

A NOTE ON MANY-SERVER QUEUEING SYSTEMS WITH ORDERED ENTRY, WITH AN APPLICATION TO CONVEYOR THEORY

W. M. NAWIJN,* *Twente University of Technology*

Abstract

Consider a many-server queueing system in which the servers are numbered. If a customer arrives when two or more servers are idle he selects the server with lowest index (this is called the ordered entry selection rule). An explicit expression for the traffic handled by the various servers in a $GI/M/s$ queueing system with ordered entry is derived. For the $M/M/s$ queueing system the probability distribution of the number of busy servers among the first $k (k = 1, 2, \cdots, s)$ servers will be given. Finally, a formula for the traffic handled by the first server in an $M/D/s$ system will be derived. All results are derived under steady-state conditions. As an application some numerical data for the server utilizations will be given and compared to data obtained from simulation studies of a closed-loop continuous belt-conveyor.

SERVER UTILIZATION; CONVEYOR THEORY

Introduction

Consider a many-server queueing system in which the servers are numbered from 1 to s. Assume that any arriving customer who finds more than one server idle, selects among the idle servers the one with lowest index. This selection rule is known as the 'ordered entry rule', see e.g. Disney [2]. In using this selection rule the traffic handled by the various servers will differ. It is the objective of this paper to derive an explicit expression for the traffic handled by the kth server $(1 \leqq k \leqq s)$ for the queueing systems $GI/M/s$ and $M/M/s$, under steady-state conditions. Moreover, for the $M/M/s$ system, the probability distribution of the number of busy servers among the servers with index $\leqq k$, will be given. Finally, as an isolated result, the traffic handled by the server with lowest index in the $M/D/s$ system will be derived.

The problem considered is of interest for the analysis of so-called closed-loop circulating conveyors in which the 'ordered entry rule' is induced by the

Received 4 December 1981.

* Department of Mechanical Engineering, Twente University of Technology, P.O. Box 217, 7500 AE Enschede, The Netherlands.

geometrical configuration of the system, see e.g. Pritsker [7], Phillips and Skeith [6] and Proctor et al. [8]. In these studies it was found that the traffic handled (i.e. server utilization) by the various servers seems to be independent of the circulation time of the conveyor (see also Nawijn and Rooda [5]). By assuming a zero circulation time the conveyor models become equivalent to many-server queueing systems with ordered entry.

1. The $GI/M/s$ queueing system

Consider the $GI/M/s$ queueing system with interarrival-time distribution function $G(\cdot)$, with mean λ^{-1}, and with service-time distribution function $1-e^{-\mu x}$, $x>0$, in which the servers are numbered from 1 to s. Let us denote by $X(t)$ the number of customers in the system at time t and let $Y^{(k)}(t)$ denote the number of busy channels among the servers numbered 1 to k. Define the imbedded Markov chain $(X_n, Y_n^{(k)}) = (X(t_n-0), Y^{(k)}(t_n-0))$, in which t_n denotes the nth arrival epoch, with state space $\{0,1,\cdots,\cdots\}\times\{0,1,\cdots,k\}$. Obviously, $Y_n^{(k)} \leqq X_n$ and $Y_n^{(k)} = k$ if $X_n \geqq s$. It will be assumed that $\lambda < \mu s$ and the servers are selected according to the ordered entry rule.

Let the stationary distribution of the Markov chain, which exists since $\lambda < \mu s$, be denoted by

$$P_{n,m}^{(k)} = \Pr\{X = n,\ Y^{(k)} = m\} = \lim_{r\to\infty} \Pr\{X_r = n,\ Y_r^{(k)} = m \mid X_1,\ Y_1^{(k)}\} \tag{1}$$

for $n \geqq 0$, $0 \leqq m \leqq \min(n,k)$.

Recall the following result (see Takács [10], Theorem 1, p. 148):

$$\Pr\{X = n\} = A\omega^{n-s}, \qquad n \geqq s, \tag{2}$$

$$A = 1\Big/\left[(1-\omega)^{-1} + \sum_{j=1}^{s}\binom{s}{j}\frac{1}{C_j(1-\Phi_j)}\left\{\frac{s(1-\Phi_j)-j}{s(1-\omega)-j}\right\}\right], \tag{3}$$

in which ω is the only root of the equation $\omega = \Phi(s\mu(1-\omega))$ in the unit circle where $\Phi(\zeta) = \int_0^\infty e^{-\zeta x}dG(x)$ and $\Phi_j = \Phi(j\mu)$ and

$$C_j = \prod_{r=1}^{j}\frac{\Phi_r}{1-\Phi_r}, \qquad j = 1,2,\cdots; \quad (C_0 \equiv 1). \tag{4}$$

Introducing the generating function

$$H^{(k)}(z_1,z_2) = \sum_{n=0}^{s-1}\sum_{m=0}^{k} P_{n,m}^{(k)} z_1^n z_2^m \tag{5}$$

it follows by standard methods that $H^{(k)}(z_1,z_2)$ satisfies the following integral equation:

$$
\begin{aligned}
H^{(k)}(z_1, z_2) &= \int_0^\infty f(x, z_1 z_2) H^{(k)}\left\{f(x, z_1), \frac{f(x, z_1 z_2)}{f(x, z_1)}\right\} dG(x) \\
&\quad + z_1(1 - z_2) \sum_{i=k}^{s-1} P_{i,k}^{(k)} \int_0^\infty e^{-\mu x} f(x, z_1)^{i-k} f(x, z_1 z_2)^k dG(x) \\
&\quad + A \int_{x=0}^\infty \int_{t=0}^x e^{s\mu\omega t}\{f(x, z_1) - 1 + e^{-\mu t}\}^{s-k}\{f(x, z_1 z_2) - 1 + e^{-\mu t}\}^k \\
&\quad \cdot s\mu dt dG(x) - A z_1^s z_2^k, \quad \text{where } f(x, z) = 1 - e^{-\mu x} + z e^{-\mu x}.
\end{aligned} \tag{6}
$$

Inserting $z_2 = 1$ one obtains an integral equation for $H^{(k)}(z_1, 1)$ solved by Takács [10], p. 151. Using the same method one can solve the integral equation for $H^{(k)}(1, z)$ obtained by taking $z_1 = 1$.

Define the binomial moments

$$
H_j^{(k)} = \frac{1}{j!} \frac{d^j}{dz^j} H^{(k)}(1, z)\Big|_{z=1}, \qquad j = 0, 1, \cdots, k. \tag{7}
$$

Observe that

$$
\begin{aligned}
H_0^{(k)} &= H^{(k)}(1, 1) = \Pr\{X \leqq s - 1\} = 1 - A/(1 - \omega) \\
H_k^{(k)} &= \sum_{i=0}^{s-1} P_{i,k}^{(k)},
\end{aligned} \tag{8}
$$

where $H_k^{(k)}$ is the probability that an arriving customer immediately enters one of the servers $k + 1, \cdots, s$.

It turns out that

$$
\frac{H_r^{(k)}}{C_r} = \sum_{j=r+1}^{k} \frac{A\binom{k}{j}}{C_j(1 - \Phi_j)} \left\{\frac{s(1 - \Phi_j) - j}{s(1 - \omega) - j}\right\} + H_k^{(k)} \sum_{j=r+1}^{k+1} \frac{\binom{k}{j-1}}{C_{j-1}} \tag{9}
$$

for $r = 0, 1, \cdots, k$. Taking $r = 0$ in (9) and using (8) we obtain

$$
H_k^{(k)} = \left[1 - \frac{A}{1 - \omega} - \sum_{j=1}^{k} \frac{A\binom{k}{j}}{C_j(1 - \Phi_j)} \left\{\frac{s(1 - \Phi_j) - j}{s(1 - \omega) - j}\right\}\right] \Big/ \sum_{j=0}^{k} \binom{k}{j} \frac{1}{C_j}. \tag{10}
$$

The distribution of $Y^{(k)}$ can be determined from the binomial moments by

$$
\Pr\{Y^{(k)} = m\} = \Pr\{Y^{(k)} = m, X \leqq s - 1\} = \sum_{r=m}^{k} (-1)^{r-m} \binom{r}{m} H_r^{(k)}, \tag{11}
$$

when $0 \leqq m \leqq k - 1$

and

$$
\Pr\{Y^{(k)} = k\} = \Pr\{X \leqq s - 1, Y^{(k)} = k\} + \Pr\{X \geqq s\} = H_k^{(k)} + \frac{A}{1 - \omega}. \tag{12}
$$

Theorem 1. Let η_k, $k = 1, 2, \cdots, s$, denotes the traffic handled by the kth server, then

$$\eta_k = \rho(H_{k-1}^{(k-1)} - H_k^{(k)}) + \frac{\rho}{s}\frac{A}{1-\omega}, \quad \text{where } \rho = \frac{\lambda}{\mu} \text{ and } H_0^{(0)} \equiv 1 - A/(1-\omega). \tag{13}$$

Proof. Observe that there are two mutually exclusive ways in which an arriving customer can enter the kth server: either the kth server is the server with lowest index among all idle servers (call this event E_1) or all servers are busy upon his arrival and he eventually will enter the kth server (call this event E_2). Now it is readily seen that $\Pr\{E_1\} = H_{k-1}^{(k-1)} - H_k^{(k)}$, $k = 1, 2, \cdots, s$, where $H_0^{(0)} = \Pr\{X \leqq s-1\}$. Since the service times are independent exponentially distributed and independent of the arrival process a waiting customer may enter any server with equal probability, hence

$$\Pr\{E_2\} = \frac{1}{s}\Pr\{X \geqq s\} = \frac{1}{s}\frac{A}{1-\omega}$$

(cf. (2)). It follows that the expected number of customers served by the kth server per unit of time is given by

$$\lambda(H_{k-1}^{(k-1)} - H_k^{(k)}) + \frac{\lambda}{s}\cdot\frac{A}{1-\omega},$$

from which η_k is found by multiplying by μ^{-1}, the expected service time.

Note that

$$\sum_{k=1}^{s}\eta_k = \rho(H_0^{(0)} - H_s^{(s)}) + \rho\cdot\frac{A}{1-\omega} = \rho\left\{1 - \frac{A}{1-\omega}\right\} + \rho\frac{A}{1-\omega} = \rho, \tag{14}$$

as indeed it must.

2. The $M/M/s$ queueing system

As a particular case suppose that the customers arrive according to a stationary Poisson process, i.e. $G(x) = 1 - e^{-\lambda x}$, $x > 0$. Then, since $\Phi(\xi) = \lambda/(\lambda + \xi)$, we have

$$\Phi_j = \rho/(\rho + j) \quad \text{and} \quad C_j = \frac{\rho^j}{j!}, \qquad j = 1, 2, \cdots \tag{15}$$

and $\omega = \rho/s$.

Substituting this into (10) yields

$$H_k^{(k)} = \frac{\dfrac{\rho^k}{k!}}{\displaystyle\sum_{j=0}^{k}\frac{\rho^j}{j!}}\Pr\{X \leqq s\} - \Pr\{X = s\}. \tag{16}$$

Hence, from Theorem 1, we obtain for $k = 1, 2, \cdots, s$

$$\eta_k = \rho\{E_{1,k-1}(\rho) - E_{1,k}(\rho)\}\Pr\{X \leqq s\} + \Pr\{X > s\}, \tag{17}$$

in which $E_{1,k}(\rho) = (\rho^k/k!)/\Sigma_{j=0}^{k} \rho^j/j!$, Erlang's loss formula, and the probabilities $\Pr\{X \leqq s\}$ and $\Pr\{X > s\}$ are well known.

Moreover it follows by straightforward calculations from (9)–(12) that

$$\begin{aligned} \Pr\{Y^{(k)} = n\} &= \frac{\dfrac{\rho^n}{n!}}{\displaystyle\sum_{j=0}^{k} \frac{\rho^j}{j!}} \Pr\{X \leqq s\}, \qquad n = 0, 1, \cdots, k-1, \\ &= \frac{\rho^k/k!}{\displaystyle\sum_{0}^{k} \rho^j/j!} \Pr\{X \leqq s\} + \Pr\{X > s\}, \qquad n = k. \end{aligned} \tag{18}$$

Note that the distribution of $Y^{(k)}$ also holds for an arbitrary point in time and moreover that the quantity η_k can also be interpreted as the utilization of the kth server, i.e. the fraction of time that the server is busy.

Remark. The relations (17) and (18) could also be obtained using the following observation. Let $p(i)$ and $q(i)$ be respectively the steady-state probabilities for the $M/M/s$ loss and queueing system with ordered entry, in which the vector $i = (i_1, i_2, \cdots, i_s)$ denotes the state of the servers, that is $i_j = 1$ when server j is busy and $= 0$ otherwise.

By comparing the steady-state balance equations for both systems when $\{X \leqq s\}$ it follows, noting that $s\mu \Pr\{X = s+1\} = \lambda \Pr\{X = s\}$, that $q(i) = cp(i)$ where c is some positive constant. In view of this property it is readily seen that

$$\eta_k = \eta_k(\text{loss}) \cdot \Pr\{X \leqq s\} + \Pr\{X > s\}, \tag{19}$$

in which $\eta_k(\text{loss})$ is the traffic handled by the kth server in the loss system.

In an $M/G/s$ loss system with ordered entry the traffic handled by the kth server easily follows from the fact that the input rate of the kth server is the difference between the overflow rates of the $(k-1)$th and kth server and is given by

$$\eta_k(\text{loss}) = \rho\{E_{1,k-1}(\rho) - E_{1,k}(\rho)\}, \qquad k = 1, 2, \cdots, s. \tag{20}$$

Moreover, observe that the first k servers in the $M/M/s$ loss system form an $M/M/k$ loss system. Hence, in view of $q(i) = cp(i)$, it follows that

$$\Pr\{Y^{(k)} = n \mid X \leqq s\} = \frac{\rho^n/n!}{\displaystyle\sum_{0}^{k} \rho^j/j!}, \qquad n \leqq k \leqq s, \tag{21}$$

from which (18) is a ready consequence.

Relations (17) and (18) have been known for a long time, see Cooper [3], p. 124, who attributes (17) to Kendrick (1923). These results can be generalized for the heterogeneous-server case using the same argument, see Cooper [4].

3. The $M/D/s$ queueing system

As the $M/G/s$ system is much harder to analyse than the $GI/M/s$ system it is questionable if similar exact results could be found for the $M/G/s$ system with ordered entry. In the particular case $M/D/s$, however, the following isolated result can be proved.

Theorem 2. In a stationary $M/D/s$ queueing system with ordered entry the traffic handled by the first server is given by

$$\eta_1 = \rho \Big/ \left[\rho + \exp\left\{ -\sum_{n=1}^{\infty} \frac{1}{n} \sum_{j=ns}^{\infty} \frac{(\rho n)^j}{j!} e^{-\rho n} \right\} \right] . \tag{22}$$

Proof. Let the customer starting a busy period of the first server be called c_1. Now observe that such a busy period is always composed of the service times of the customers $c_1, c_{s+1}, c_{2s+1}, \cdots$, as long as they arrive during this busy period initiated by c_1, provided of course that the ordered entry rule is applied. Consequently, this busy period is equivalent to a busy period in an $E_s/D/1$ queueing system, i.e. the interarrival-time distribution is the s-fold convolution of the negative exponential distribution $1 - e^{-\lambda x}$, $x \geqq 0$. The above observation originated from Pollaczek's derivation of the waiting-time distribution in the $M/D/s$ queueing system, see e.g. Riordan [9], p. 117.

Let $E\{p\}$ denote the expected length of a busy period in a $GI/G/1$ queueing system. Then, see e.g. Cohen [1], p. 286,

$$E\{p\} = \frac{1}{\mu} \exp\left\{ \sum_{n=1}^{\infty} \frac{1}{n} . \Pr(s_n > 0) \right\} , \tag{23}$$

where

$$s_n = \sum_{i=1}^{n} (\tau_i - \sigma_{i+1}), \qquad n = 1, 2, \cdots ,$$

in which τ_i denotes the service time of the ith arriving customer and σ_{i+1} denotes the interarrival time between customer i and $i+1$.

Note that for the system $E_s/D/1$ we have (i) $\tau_i = \mu^{-1}$ $(i = 1, 2, \cdots)$ and (ii) σ_{i+1} is the sum of s independent negative exponentially distributed variables with mean λ^{-1}.

Hence

$$\Pr\{s_n > 0\} = 1 - \sum_{j=0}^{ns-1} \frac{(n\rho)^j}{j!} e^{-n\rho} = \sum_{j=ns}^{\infty} \frac{(n\rho)^j}{j!} e^{-n\rho}. \tag{24}$$

Observe that the length of an idle period of the first server is negative exponentially distributed with mean λ^{-1} and moreover that the sequence of idle and busy periods are mutually independent.

Let us denote the state of the server at time t by ε_t where $\varepsilon_t = 1$ if the server is busy at time t and $\varepsilon_t = 0$ otherwise. The process $\{\varepsilon_t, t > 0\}$ is an alternating renewal process, which is obviously regenerative with respect to the sequence of busy cycles, hence

$$\eta_1 = \lim_{t\to\infty} \frac{1}{t} \int_0^t \varepsilon_x dx = \lim_{t\to\infty} \Pr\{\varepsilon_t = 1\} = \frac{(1/\mu)E\{p\}}{(1/\lambda)+(1/\mu)E\{p\}} \quad \text{(a.s.)}. \tag{25}$$

The theorem now immediately follows from (23), (24) and (25).

Remark. Relation (22) could also be written as

$$\eta_1 = \rho/(1+\rho-P_{\text{delay}}) \tag{26}$$

in which P_{delay} gives the stationary delay probability in an $M/D/s$ queueing system, see Riordan [9], p. 117.

4. An application to a closed-loop continuous belt-conveyor

Consider a conveyor belt moving continuously in a closed loop with uniform speed. At some point along the conveyor jobs, arriving according to some stochastic point process, are loaded onto the conveyor. Each job has to be processed at one of a number of identical workstations, situated along the conveyor. Since there are no buffer storages at the stations, a server (i.e. an operator) has to wait for an arrival of a job on the conveyor once he has finished the processing of a job. A job that has not been unloaded by one of the servers will recirculate. The time, T say, for the conveyor to make one revolution is called the recirculation time. This conveyor model has been studied, using simulation, by several authors, see [5]–[8]. One of the results found was that the server utilizations seem to be independent of the recirculation time. This insensitivity property suggests choosing $T = 0$ and solving the resulting model analytically. Clearly when $T = 0$ the model becomes equivalent to a many-server queueing system with ordered entry. Notice that for $T > 0$ the conveyor model can be viewed as a congestion system with repeated calls, in which the repetition time is constant, a model known from telephone traffic theory, see e.g. Riordan [9], p. 94.

We illustrate below the use of the formulas (13) and (17) and compare some numerical results with simulation results obtained by Pritsker [7] and Nawijn and Rooda [5]. Before doing so it is worth mentioning that the server utilizations for large values of T can be approximated using a classic method from telephone traffic theory, see Wilkinson [11], p. 426. The idea of this method is to replace the

system under consideration by an 'equivalent' $M/M/s$ loss system with arrival rate $\lambda + \varepsilon$, in which ε measures the overflow rate in the original system. Now ε is determined such that the traffic handled by the s servers in this loss system equals $\rho = \lambda/\mu$, where μ^{-1} is the mean processing time. Hence ε is the (unique) positive root of

$$\varepsilon = (\lambda + \varepsilon)E_{1,s}(\rho + \varepsilon/\mu) \tag{27}$$

and the server utilizations are determined from (20) with ρ replaced by $\rho + \varepsilon/\mu$.

Tables 1 and 2 give the server utilizations when $s = 5$ for a Poisson and a deterministic arrival process. The simulation results from [5] and [7] in these tables show the insensitivity property. The results are in reasonable agreement with the numerical values obtained by the analytical formulas. There is a tendency for the utilizations of the lower-indexed servers to decrease and for the higher-indexed to increase as T increases.

TABLE 1
Server utilizations in $M/M/s$ conveyor system ($\lambda = 1$)

T	ρ	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5	source
0	4	0.889	0.854	0.810	0.755	0.691	(17)
1	4.023	0.889	0.854	0.814	0.763	0.703	[5]
5	3.999	0.874	0.846	0.805	0.761	0.713	[5]
20	4.002	0.869	0.841	0.807	0.765	0.720	[5]
∞	4	0.872	0.844	0.809	0.764	0.709	(20), (27)

TABLE 2
Server utilizations in a $D/M/5$ conveyor system ($\lambda = 1$)

T	ρ	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5	source
0	2	0.787	0.628	0.390	0.158	0.036	(13)
2	2.030	0.779	0.639	0.384	0.170	0.043	[7]
4	2.068	0.792	0.634	0.391	0.177	0.045	[7]

Remark. The simulation results of Pritsker [7] are based on a run length of 4000 time units ($\lambda = 1$). The results from [5] are based on regenerative cycles giving 95 per cent confidence intervals ranging from (0.875, 0.899) and (0.664, 0.740) for ρ_1 and ρ_5 respectively at $T = 1$ to (0.867, 0.871) and (0.715, 0.725) for ρ_1 and ρ_5 respectively at $T = 20$.

Table 3 gives as a typical example the effect of the interarrival-time distribution on the server utilizations in the $GI/M/5$ system with ordered entry for $\rho = 4$. Apparently the server utilizations of the lower-indexed servers increase as the variance of the interarrival-time distribution decreases. The lower the value of ρ/s the stronger this effect will be; see Table 4 with $s = 2$ and $\rho = 1$.

TABLE 3

G	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5
M	0.889	0.854	0.810	0.755	0.691
E_2	0.902	0.866	0.816	0.749	0.667
E_4	0.911	0.874	0.821	0.745	0.649
E_{10}	0.916	0.880	0.825	0.743	0.636
D	0.921	0.884	0.828	0.742	0.625

TABLE 4

G	ρ_1	ρ_2
M	0.583	0.417
E_2	0.610	0.390
E_4	0.629	0.371
E_{10}	0.644	0.356
D	0.655	0.345

References

[1] COHEN, J. W. (1969) *The Single Server Queue*. North-Holland, Amsterdam.

[2] DISNEY, R. L. (1963) Some results of multichannel queueing problems with ordered entry — an application to conveyor theory. *J. Industrial Engineering* **14**, 105–108.

[3] COOPER, R. B. (1972) *Introduction to Queueing Theory*. Macmillan, New York.

[4] COOPER, R. B. (1976) Queues with ordered servers that work at different rates. *Opsearch* **13**, 69–78.

[5] NAWIJN, W. M. AND EELKMAN ROODA, J. (1980) An analysis of operating characteristics of closed-loop continuous conveyors, using simulation and queueing approximations. *Mechanical Comm.* **6**, Twente University of Technology, Enschede, The Netherlands.

[6] PHILLIPS, D. T. AND SKEITH, R. W. (1969) Ordered entry queueing networks with multiple services and multiple queues. *AIIE Trans.* **1**, 333–342.

[7] PRITSKER, A. A. B. (1966) Applications of multichannel queueing results to the analysis of conveyor systems. *J. Industrial Engineering* **17**, 14–21.

[8] PROCTOR, C. L., EL SAYED, E. A. AND ELAYAT, H. A. (1977) A conveyor system with homogeneous and heterogeneous servers with dual input. *Internat. J. Production Res.* **15**, 73–85.

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

[10] TAKÁCS, L. (1962) *Introduction to the Theory of Queues*. Oxford University Press, New York.

[11] WILKINSON, R. I. (1956) Theories for toll traffic engineering in the U.S.A. *Bell System Tech. J.* **35**, 796–802.