Queues with Ordered Servers that Work at Different Rates: An Exact Analysis of a Model Solved Approximately by Others

Robert B. Cooper
Department of Computer and Information Systems, Florida Atlantic University, P.O. Box 3091, Boca Raton, FL 33431-0991, U.S.A.

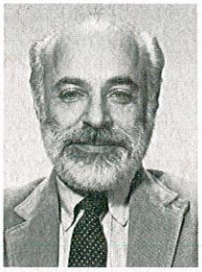
Received 15 August 1986
Revised 26 November 1986

Keywords: Queue, Heterogeneous Server, Ordered Server, Ordered Entry.

In a recent paper, Ibe and Maruyama [3, p. 16] consider the following model of a heterogeneous multiserver queueing system, which they describe in Section 1 of their paper: "There are n communication links numbered $1, 2, \ldots, n$, which are accessed by arriving messages in a fixed order. Specifically, an arriving message is processed by the lowest-numbered idle link, if such a link exists. When all links are busy, a single queue is formed and messages enter for service on a first-come first-served basis. We assume that messages arrive in a Poisson manner with rate λ, that message lengths are exponentially distributed with mean $1/\mu$, and that $C_{i-1} \geqslant C_i$, where C_i is the capacity of link i, $i = 2, \ldots, n$. We want to compute the utilization factor of each link and the expected message delay."

The authors assert that an exact analysis requires the solution of 2^n linear equations, which is impractical except when n is small. Therefore, they propose an approximation, which they compare with the exact results for $n = 2$ and $n = 3$, and with simulation results for $n = 4$, 5, and 10.

The purpose of this short note is to point out that the model of Ibe and Maruyama is a special case of the one considered by the present author [1], who gives an exact and easily computable solution for the case of arbitrary n, and which remains valid under less restrictive assumptions about the arrival process and the queue discipline. More specifically, Cooper's model and its solution [1, pp. 72–73] are stated as follows.



Robert B. Cooper received his B.S. degree in Science from Stevens Institute of Technology in 1961, his M.S. degree in Systems Engineering and Operations Research in 1962, and his Ph.D. in Electrical Engineering in 1968, both from the University of Pennsylvania. He has been associated with AT&T Bell Laboratories, Georgia Tech, The University of Michigan, and The New Mexico Institute of Mining and Technology, and is presently a professor at Florida Atlantic University. His main research interest is queueing theory.

North-Holland

"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 $\mu_k = \mu(1)x_1 + \cdots + \mu(s)x_s$ 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 μ_k $(k \leqslant s)$ not only depends on the number k of busy servers, but also on the identities of the busy servers.)

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

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

and let

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

Finally, we define

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

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

$$P_{s+i} = [B_s/(1 + AB_s)]\,A_i, \quad i = 0, 1, \ldots. \tag{6}$$

(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

$$p_j = \frac{\lambda}{\mu(j)}(B_{j-1} - B_j)\left(1 - \sum_{i=1}^{\infty} P_{s+i}\right) + \sum_{i=1}^{\infty} P_{s+i}, \quad j = 1, 2, \ldots, s, \ B_0 = 1." \tag{7}$$

To see the correspondence between the two models, note that the quantities denoted by Ibe and Maruyama as n (number of links), μC_i (service rate of link i), and P_i (utilization of link i) are in Cooper's notation s, $\mu(i)$ and p_i, respectively. To specialize Cooper's model to agree with Ibe and Maruyama's, take $\lambda_k = \lambda$ and $\mu_k = C\mu$ for $k = s, s+1, \ldots$. Then,

$$A_i = (\lambda/C\mu)^i = \rho^i, \quad \rho < 1, \ i = 1, 2, \ldots,$$

and thus

$$A = \rho/(1 - \rho) \quad \text{and} \quad \sum_{i=1}^{\infty} P_{s+i} = \rho B_s/(1 - \rho + \rho B_s).$$

Then, equation (7) above gives *exactly* what equation (20) of Ibe and Maruyama gives *approximately*, and

Also, Ibe and Maruyama approximate the expected service time and the expected waiting time (from arrival to start of service), which they denote by T and W, respectively. In Cooper's notation, these quantities are given, exactly, by

$$T = \frac{1}{\lambda}(p_1 + \cdots + p_s) \quad \text{and} \quad W = \frac{1}{\lambda}\sum_{i=1}^{\infty} iP_{s+i} = \frac{1}{\lambda(1-\rho)}\frac{\rho B_s}{1-\rho+\rho B_s}.$$

The corresponding approximations are given by Ibe and Maruyama's equations (23) and (24).

The exact results are all as easily computed as the approximations, the only nontrivial calculation being that of the quantities B_i $(i = 1, 2, \ldots, s)$, defined by (3) and calculated iteratively according to (2). (For an illustration of the iterative calculation of (2), see [2].)

References

[1] R.B. Cooper, Queues with ordered servers that work at different rates, *Opsearch* **13** (2) (1976) 69–78.

[2] R.B. Cooper and M.K. Solomon, The average time until bucket overflow, *ACM Trans. Database Systems* **9** (3) (1984) 392–408.

[3] O.C. Ibe and K. Maruyama, An approximation method for a class of queueing systems, *Performance Evaluation* **5** (1) (1985) 15–27.