韓國經營科學會誌
第24卷 第2號
1999年 6月

An Approximate Analysis of the Queueing Systems with Two Deterministic Heterogeneous Servers

Jeongseob Kim*

Abstract

A new approximation method for finding the steady-state probabilities of the number of customers present in queueing systems with Poisson arrivals and two servers with different deterministic service times with infinite waiting room capacity is developed. The major assumption made for the approximation is that the residual service times of the servers have mutually independent uniform distributions with densities equal to the reciprocals of the respective service times. The method reflects the heterogeneity of the servers only through the ratio of their service times, irrespective of the actual magnitudes and difference. The transition probability matrix is established and the steady-state probabilities are found for a variety of traffic intensities and ratios of the two service times; also the mean number of customers present in the system and in the queue, and server utilizations are found and tabulated. The method was validated by simulation and turned out to be very sharp.

1. Introduction

Queueing systems with heterogeneous servers are very difficult to analyze. Even though there are some exceptions, studies with some closed form expressions for the usual performance measures are mostly limited to the systems with Poission arrivals and **exponential** servers. In this paper we present a simple approximation method to compute the steady-state probabilities of the number of customers present in queueing systems with Poisson arrivals, two servers with different **deterministic** service times, and waiting room of infinite capacity. In what follows, the subscript *i* to the server description in the usual notation of queueing system, for instance, $M/M_i/s$, indicates that the service rates of the servers are not necessarily identical.

Early works on heterogeneous server queue focus on the impact of heterogeneity on the performance of the system. Gumbel [4] was the

* Department of Business Administration, Taegu University, Korea

first to consider a queueing system with heterogeneous servers. Gumbel obtained expressions for the usual expected measures for $M/M_i/s$ system where customers select randomly any idle server. A nondimensional ratio, $\sqrt{\sum_{i=1}^{s}(\mu_i-\bar{\mu})^2/\sum_{i=1}^{s}\mu_i}$, where μ_i is mean service rate of server i, is used as a measure of heterogeneity in analyzing the error incurred in assigning each server the arithmetic mean $\bar{\mu}$ with expected number in the system as a criterion. It is shown that the larger the heterogeneity of the servers (even at constant mean rate) the larger the waiting line. It is also seen that the error is quite small for a fair extent of variability among the servers. Ancker and Gafarian [1] investigated the problem posed by Gumbel [4] further. They set an upper limit on the queue size and allowed customers to renege while waiting for service. Various steady-state results were obtained.

Several authors considered systems with different rules of assigning arriving customers of *single* class to different servers. Lemoine [9] considered a queue with heterogeneous servers in which arriving customers are assigned to servers according to an autonomous Markov chain, which is a generalization of random server selection. Establishing stability conditions, Lemoine derived stationary distributions for waiting times and the system response times of successive customers. He also obtained limits for the expected utilization and the expected number of customers in the system. Cooper [2] considered an $M/M_i/s$ queue with arbitrarily numbered servers, in which customers arrive according to a state dependent Poisson process and are served by the lowest-numbered idle server. This generalization of server selection encompasses several arrival processes and queue disciplines: for examples, balking, finite waiting room. He obtained explicit forms to compute the utilization of each server and the probability distribution of the number of customers waiting in the queue, without explicitly solving the system of balance equations which usually arises.

The problem of assigning *multiple* class of customers to heterogeneous servers was also considered by some authors (Winston [15,16,17], Derman et al. [3]). Winston [15,16] considered optimal rules for assigning several classes of customers to heterogeneous servers which minimize a discounted reward of a continuous time Markov decision process superimposed on the queueing system, in which the distribution of a customer's service time is dependent on both the class of the customer and the type of server to which he/she is assigned. He presented conditions that ensure that the discounted number of service completions is maximized by assigning customers with longer service time to faster servers. Winston [17] modified the assumption about the distribution of service times in his earlier work [15]; the distribution of a customer's service time depends only his/her class. It is shown that the long-term expected reward earned over an infinite horizon depends on a single critical number.

Heterogeneity among servers raises control issues to optimize some performance measures, usually to minimize some cost function. The main ideas are to utilize the faster servers as efficiently as possible using slower servers as aids, to determine the service capacities of the servers to optimize some performance measures, and to decide some policy for operating the queueing system. Several authors studied the so-called "threshold" service discipline under which if the number of customers waiting for service is less than *m*, the slow server remains idle, and only is invoked if the threshold is surpassed, in which case the customer in the (m+1)-th position in the queue goes to the slow server (Krishnamoorti [6], Larsen [7], Larsen and Agrawala [8], Lin and Kumar [10], Iliadis and Lien [5]). Larsen [7] and Larsen and Agrawala [8] conjectured that the optimal queue discipline that minimizes the sojourn time in an $M/M_i/2$ queue is of threshold type, and analyzed the performance of this discipline in detail. It is shown that the higher the ratio of the service rates, the more improvement is provided through threshold discipline. No significant improvement is noted until this ratio exceeds two, however. Lin and Kumar [10] provided a formal proof for Larsen's conjecture.

Some authors studied the optimization of service rates with some objective function. Singh [11] studied $M/M_i/2$ with ordered servers and balking. Singh derived exact closed form solutions to determine the rate of the slow server to minimize the average queue length, average number in system. He also obtained a necessary and sufficient condition for the heterogeneous server system to be better than the corresponding homogeneous system with service rate of arithmetic mean. Singh [12] later generalized his work to $M/M_i/3$. In a multi-server Markovian queue with no waiting room and with ordered selection of servers, Tahara and Nishida [13] found that the optimal service rates of each server which minimize the rate of lost customers are positive and different for each server.

Although the above (limited) survey implies certain amount of research on queueing systems with heterogeneous servers, no study has been reported for queueing systems with **deterministic** service time. In this paper (in Section 2) we present a simple discretization method to approximately compute the steady-state probabilities of the number of customers present in queueing systems with Poisson arrivals, two servers with different deterministic service times, and a waiting room of infinite capacity. In doing so, we first define various notations and the states of the involved stochastic system, and then set up the usual balance equations for Markov Chains. The simultaneous equations are solved by an iterative method and the results are compared with those from simulation to validate our approximation method.

2. The queue $M/D_i/2$

We consider a queue with Poisson arrival process with parameter λ, two servers with heterogeneous deterministic service times D_1 and D_2, $(D_1 \leq D_2)$, time units, and a waiting room of infinite capacity. Upon arriving if a customer finds both servers busy, he/she joins the common queue; if only one server is available he/she selects that server; if both servers are free he/she select server 1, the faster server.

The deterministic service times lead us to the following observations.

- Since the service times are constant, any customer in service at server *i* at some time *t* will have left the system at time $t+D_i$.

- The customers present at time $t{+}D_l$ are those customers who (1) were waiting in the queue at time t, or (2) were under service at time t by server 2, or (3) arrived during $(t,t{+}D_l]$.

These observations suggest a way of approximating the behavior of the system-discretizing the continuous system by sampling every D_l units of time. Let $\mathbf{x}(t)=(i,j,k)$ denote the state of the system at time t, where i and j denote the state of servers 1 and 2, respectively. These indices will take a value of 1 if the corresponding server is busy, 0 otherwise. The third index k represents the number of customers waiting in the queue. Then we get the state space $S=\{(i,j,k) \mid (0,0,0),\ (0,1,0),\ (1,0,0),\ (1,1,n),\ n \geq 0\}$ and possible transitions from time t to $t{+}D_1$ as following:

- $(0,0,0) \rightarrow \mathbf{x} \in S \setminus \{(0,1,0)\}$,
- $(1,0,0),\ (0,1,0),\ (1,1,0) \rightarrow \mathbf{x} \in S$,
- $(1,1,1) \rightarrow \mathbf{x} \in S \setminus \{(0,0,0)\}$,
- $(1,1,n) \rightarrow (1,1,m),\ m=n-2,\ n-1,\ n,\ n+1, \ldots;\ 2 \leq n.$

Let $Y_1(t)$ and $Y_2(t)$, $(Y_1(t)>0,\ Y_2(t)>0)$, be the random variables representing the residual service times of server 1 and 2, respectively, at an arbitrary sampling point t if the corresponding server is busy at that moment. We assume these two random variables are mutually independent and have uniform distribution with density $1/D_1$ and $1/D_2$, respectively. Since the service times are deterministic, these assumed distributions tend to spread wider than actual. Verifying this assumption is not trivial partly because the servers are not always busy so that the underlying stochastic process is not of a renewal type. But assuming renewal process, it can be easily shown that the residual service time is a uniform random variable by

$$\begin{aligned} B(t) &= \frac{1}{D}\int_0^t (1-\delta(x-D))dx \\ &= \begin{cases} t/D & \text{if} \quad t \leq D \\ 1 & \text{if} \quad D \leq t, \end{cases} \end{aligned} \tag{1}$$

where

$B(t)$ = Cumulative distribution function of the residual service time,

$$\delta(x) = \begin{cases} 0 & \text{if} \quad x < 0 \\ 1 & \text{if} \quad x \geq 0 \end{cases}$$

To establish the system of balance equations of transitions among the states, we enumerate by careful probabilistic reflection, for each state $\mathbf{x}(t+D_1)$ at time $t+D_1$, all possible source states $\mathbf{x}(t)$ at time t and the corresponding events which make the transitions as shown in Table 1. In our problem the events are the residual service times of servers at time t and the number of customers arriving during $(t,\ t+D_1]$. In Table 1 the notation $n \in T$ denotes the event that n customers arrive during a time interval T. For example, $0 \in (t,\ t+Y_1(t)]$ denotes that when server 1 is busy at time t no new customers arrive from time t until it finishes the work. In particular when the interval T is $(t,\ t+D_1]$, the event $j \in (t,\ t+Y_1(t)]$ is simplified by E_j for notational convenience.

Taking some transitions into state $\mathbf{x}(t+D_1)=(1,1,0)$ as examples, we explain the transitions enumerated in Table 1. For the transitions into $\mathbf{x}(t+D_1)=(1,1,0)$, the residual service time of the servers, if they are busy at time t, and the number of customers arriving in the interval $(t,\ t+D_1]$ determine the transitions. The possible originating states at time t are (0,0,0), (1,0,0), (0,1,0), (1,1,0), (1,1,1), and (1,1,2). Let us take three origin states from these six states.

〈Table 1〉 The state transitions and the corresponding events

$\mathbf{x}(t+D_1)$	$\mathbf{x}(t)$	Events
(0,0,0)	(0,0,0)	E_0
	(1,0,0)	E_0
	(0,1,0)	$Y_2(t) < D_1 \cap E_0$
	(1,1,0)	$Y_2(t) < D_1 \cap E_0$
(1,0,0)	(0,0,0)	E_1
	(1,0,0)	$0 \in (t,\ t+Y_1(t)] \cap 1 \in (t+Y_1(t),\ t+D_1]$
	(0,1,0)	$Y_2(t) < D_1 \cap E_1$
	(1,1,0)	$\{Y_1(t)<Y_2(t) < D_1 \cap E_1\}$ $\cup \{Y_2(t)<Y_1(t) \cap 0 \in (t,\ t+Y_1(t)]$ $\cap 1 \in (t+Y_1(t),\ t+D_1]\}$
	(1,1,1)	$Y_1(t)<Y_2(t) < D_1 \cap E_0$
(0,1,0)	(1,0,0)	$1 \in (t,\ t+Y_1(t)] \cap 0 \in (t+Y_1(t),\ t+D_1]$
	(0,1,0)	$Y_2(t) > D_1 \cap E_0$
	(1,1,0)	$\{Y_2(t)>D_1 \cap E_0\} \cup \{Y_2(t)<Y_1(t) \cap 1$ $\in (t,\ t+Y_1(t)] \cap 0 \in (t+Y_1(t),\ t+D_1]\}$
	(1,1,1)	$Y_2(t) < D_1 \cap E_0$
(1,1,n), $n \geq 0$	(0,0,0)	E_{n+2}
	(1,0,0)	E_{n+2}
	(0,1,0)	$Y_2(t) < D_1 \cap E_{n-2}$
	(1,1,k), $0 \leq k \leq n+1$	$\{Y_2(t) < D_1 \cap E_{n+2-k}\} \cup \{Y_2(t) > D_1$ $\cap E_{n+1-k}\}$
	(1,1,n+2)	$Y_2(t) < D_1 \cap E_0$

- From state $\mathbf{x}(t)$=(1,0,0). Since server 1 will finish his/her current work before $t+D_1$, the residual service time of server 1 is irrelevant. If exactly two customers arrive during $(t,\ t+D_1]$, we will find the two servers busy and the waiting room empty at time $t+D_1$.
- From state $\mathbf{x}(t)$=(0,1,0). If server 2 does not finish his/her current work until $t+D_1$ and one new customer arrives during $(t,\ t+D_1]$, the state $\mathbf{x}(t+D_1)$ =(1,1,0) will be realized. If server 2 does finish his/her current work before $t+D_1$ and two new customers arrive during $(t,\ t+D_1]$, the state $\mathbf{x}(t+D_1)$ =(1,1,0) will be realized.
- From state $\mathbf{x}$(t)=(1,1,2). Only if server 2 finish his/her current work before $t+D_1$ and there is no new customer arriving during $(t,\ t+D_1]$, the state (1,1,0) will be realized.

In addition to the notations defined so far, the following ones are also used in this paper:

- $p_{\mathbf{x}}(t)$ = the probability that the system is in state $\mathbf{x} \in S$ at time t
- $\pi_{\mathbf{x}}$ = the steady state probability that the system is in state $\mathbf{x} \in S$
- $\rho = \dfrac{1}{1/D_1 + 1/D_2}$ = a given traffic intensity
- $\alpha = D_1/\mathrm{D}_2$
- $\beta = (1+\alpha)\rho$.

Now, we write the corresponding balance equation between the state $\mathbf{x}(t+D_1)$=(1,1,0) and the above mentioned source states as follows:

$$
\begin{aligned}
&P_2(t+D_1) \\
&= \Pr\{(\mathbf{x}(t)=(0,0,0)) \cap E_2\} + \Pr\{(\mathbf{x}(t)=(1,0,0)) \cap E_2\} \\
&\quad + \Pr\{(\mathbf{x}(t)=(0,1,0)) \cap (Y_2(t) > D_1) \cap E_1\} \\
&\quad + \Pr\{(\mathbf{x}(t)=(0,1,0)) \cap (Y_2(t) < D_1) \cap E_2\} \\
&\quad + \Pr\{(\mathbf{x}(t)=(1,1,0)) \cap (Y_2(t) > D_1) \cap E_1\} \\
&\quad + \Pr\{(\mathbf{x}(t)=(1,1,0)) \cap (Y_2(t) < D_1) \cap E_2\} \\
&\quad + \Pr\{(\mathbf{x}(t)=(1,1,1)) \cap (Y_2(t) > D_1) \cap E_0\} \\
&\quad + \Pr\{(\mathbf{x}(t)=(1,1,1)) \cap (Y_2(t) < D_1) \cap E_1\} \\
&\quad + \Pr\{(\mathbf{x}(t)=(1,1,1)) \cap (Y_2(t) < D_1) \cap E_0\} \\
&= \Pr\{E_2\} p_{000}(t) + \Pr\{E_2\} p_{100}(t) \\
&\quad + [\Pr\{Y_2(t) > D_1\}\Pr\{E_1\} + \Pr\{Y_2(t) < D_1\}\Pr\{E_2\}] \\
&\qquad (p_{010}(t) + p_{110}(t)) \\
&\quad + [\Pr\{Y_2(t) > D_1\}\Pr\{E_0\} + \Pr\{Y_2(t) < D_1\}\Pr\{E_1\}] \\
&\qquad p_{111}(t) \\
&\quad + \Pr\{Y_2(t) < D_1\}\Pr\{E_0\}]\ p_{112}(t), \qquad (2)
\end{aligned}
$$

where the probabilities of the involved events can be replaced by the following relations:

$$\begin{aligned}
&\Pr\{E_j\} &&= e^{-\lambda D_1}(\lambda D_1)^j/j! &&= e^{-\beta}\beta^j/j! \\
&\Pr\{Y_2(t) > D_1\} &&= (D_2 - D_1)/D_2 &&= 1-\alpha \\
&\Pr\{Y_2(t) < D_1\} &&= D_2/D_2 &&= \alpha
\end{aligned}$$

Now, by letting $t \to \infty$ in equation (2), we get the following balance equation for the steady state probabilities.

$$\begin{aligned}
\pi_{110} = {} & \frac{e^{-\beta}\beta^2}{2}(\pi_{000}+\pi_{100}) \\
& + \frac{\alpha\beta + 2(1-\alpha)}{2} e^{-\beta}\beta(\pi_{010}+\pi_{110}) \\
& + (\alpha\beta + 1 - \alpha)e^{-\beta}\pi_{111} + \alpha e^{-\beta}\pi_{112} \qquad (3)
\end{aligned}$$

Applying the above procedure to all other states, we get the following system of balance equations for the steady state probabilities.

$$\pi_{000} = e^{-\beta}\pi_{000} + e^{-\beta}\pi_{100} + \alpha e^{-\beta}\pi_{010} + \alpha e^{-\beta}\pi_{110} \qquad (4)$$

$$\begin{aligned}
\pi_{100} = {} & \beta e^{-\beta}\pi_{000} + (\beta e^{-\beta}/2)\pi_{100} + \alpha\beta e^{-\beta}\pi_{010} \\
& + (2\alpha\beta e^{-\beta}/3)\pi_{110} + (\alpha e^{-\beta}/2)\pi_{111} \qquad (5)
\end{aligned}$$

$$\begin{aligned}
\pi_{010} = {} & (\beta e^{-\beta}/2)\pi_{100} + (1-\alpha)e^{-\beta}\pi_{010} \\
& + (1-\alpha+\alpha\beta/3)e^{-\beta}\pi_{110} + (\alpha e^{-\beta}/2)\pi_{111} \qquad (6)
\end{aligned}$$

$$\begin{aligned}
\underset{n\geq 0}{\pi_{11n}} = {} & \frac{\beta^{n+2}e^{-\beta}}{(n+2)!}\pi_{000} + \frac{\beta^{n+2}e^{-\beta}}{(n+2)!}\pi_{100} \\
& + \frac{\alpha\beta + (n+2)(1-\alpha)}{(n+2)!}\beta^{n+1}e^{-\beta}\pi_{010} \\
& + \sum_{k=0}^{n+1}\frac{\alpha\beta + (n+2-k)(1-\alpha)}{(n+2-k)!} \\
& \beta^{n+1-k}e^{-\beta}\pi_{11k} + \alpha e^{-\beta}\pi_{11(n+2)} \qquad (7)
\end{aligned}$$

Notice that the equations (4)~(7) are free of D_1 and D_2 implying that, for a given traffic intensity ρ, the heterogeneity in service times is relevant only through their ratio, irrespective of the actual magnitudes and difference. This provides generality of this approach to $M/D_i/2$.

In particular, if $D_1 = D_2 = D$, it can easily be shown that equations (4)~(7) reduce to

$$\begin{aligned}
\pi_j = {} & \frac{e^{-\lambda D}(\lambda D)^j}{j!}\sum_{k=0}^{2}\pi_k \\
& + \sum_{k=3}^{j+1}\frac{e^{-\lambda D}(\lambda D)^{j-k+2}}{(j+2-k)!}\pi_k, \quad 0 \leq j \qquad (8)
\end{aligned}$$

where π_j denotes the steady state probability that there are j customers in the system. This conforms with the general form of balance equation of *M/D/2* given in Tijms [14].

To solve this system of linear equations for a specific (ρ, α), we truncate the number in system at some integer K for which $\sum_{j=K}^{\infty}\pi_j^{\exp}$ is less than some small number ε (for example $\varepsilon = 10^{-8}$), where $\pi_j^{\exp}$ denotes the steady state probability that j customers are in the system in the *M/M$_i$/2* queue, which can be easily obtained using the usual birth-and-death process analysis. This truncation is based on the inequality,

$$\sum_{j=K}^{\infty}\pi_j^{det} \leq \sum_{j=K}^{\infty}\pi_j^{\exp}, \quad 1 \leq K \qquad (9)$$

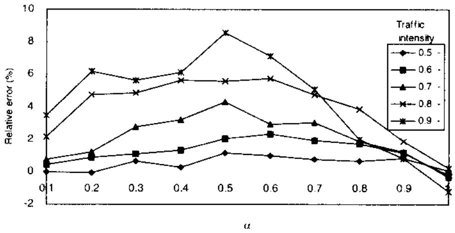
where π_j^{det} denotes the steady state probability that j customers are in the system in the *M/D$_i$/2* queue, which is intuitively reasonable by noting that the *M/D$_i$/2* queue involves less variability than the *M/M$_i$/2* queue. The resulting finite system of linear equations can be solved effectively by some iterative methods, for example, Gauss-Seidel method or successive overrelaxation method, where the probabilities $\pi_j^{\exp}$ provide a good starting point for these algorithms.

To validate our approach, we performed simulations for various (ρ, α). For each (ρ, α), we have run a simulation for 220,000 arrivals of which the first 20000 were discarded as warmup and the rest were batched to 10 samples of equal size. Table 2 compares the results from the two methods. In Table 2, Pr[Delay] denotes the probability that an arriving customer finds both servers are busy; E[L_s] the mean number of customers in the system; ρ_1 and ρ_2 the utilization of server 1 and 2, respectively. The columns under "DIS" and "SIM" are the results from discretization and simulation, respectively. For E[L_s], which is often an important performance measure for service and manufacturing systems, Table 2 also contains the standard deviations σ_{Ls} of the mean number of customers in the system obtained from the samples

〈Table 2〉 Validation of the discretization against simulation

		Pr[Delay]		E[L_s]				ρ_1		ρ_2	
ρ	α	DIS	SIM	DIS	SIM	σ_{Ls}	Err(%)	DIS	SIM	DIS	SIM
0.5	0.1	0.380	0.381	1.369	1.369	0.018	-0.03	0.481	0.479	0.686	0.694
	0.2	0.339	0.341	1.240	1.241	0.020	-0.05	0.487	0.485	0.566	0.576
	0.3	0.322	0.321	1.187	1.179	0.017	0.67	0.498	0.494	0.505	0.515
	0.4	0.315	0.316	1.165	1.162	0.020	0.25	0.512	0.509	0.470	0.482
	0.5	0.312	0.311	1.157	1.144	0.018	1.15	0.526	0.524	0.448	0.452
	0.6	0.313	0.312	1.157	1.145	0.017	0.99	0.54	0.534	0.434	0.443
	0.7	0.314	0.314	1.160	1.152	0.020	0.75	0.553	0.548	0.425	0.433
	0.8	0.317	0.316	1.165	1.158	0.020	0.65	0.565	0.562	0.419	0.425
	0.9	0.320	0.318	1.171	1.162	0.015	0.82	0.577	0.576	0.414	0.415
	1.0	0.323	0.323	1.177	1.176	0.018	0.06	0.588	0.596	0.412	0.404
0.6	0.1	0.500	0.501	1.753	1.745	0.029	0.46	0.583	0.580	0.771	0.781
	0.2	0.460	0.462	1.640	1.626	0.031	0.89	0.586	0.583	0.670	0.682
	0.3	0.442	0.445	1.591	1.574	0.030	1.07	0.596	0.592	0.615	0.63
	0.4	0.434	0.436	1.569	1.548	0.032	1.32	0.607	0.603	0.582	0.598
	0.5	0.431	0.430	1.560	1.529	0.036	2.02	0.619	0.617	0.562	0.569
	0.6	0.430	0.428	1.556	1.521	0.027	2.33	0.631	0.623	0.548	0.558
	0.7	0.431	0.430	1.555	1.526	0.031	1.91	0.642	0.635	0.539	0.549
	0.8	0.433	0.431	1.555	1.529	0.029	1.72	0.653	0.646	0.533	0.540
	0.9	0.436	0.434	1.554	1.536	0.030	1.17	0.663	0.660	0.530	0.532
	1.0	0.439	0.440	1.552	1.556	0.033	-0.25	0.673	0.682	0.527	0.521
0.7	0.1	0.623	0.626	2.304	2.286	0.054	0.77	0.686	0.684	0.841	0.853
	0.2	0.588	0.594	2.219	2.192	0.060	1.23	0.688	0.687	0.762	0.780
	0.3	0.572	0.575	2.181	2.123	0.060	2.75	0.695	0.690	0.718	0.735
	0.4	0.564	0.566	2.164	2.097	0.059	3.19	0.704	0.699	0.691	0.707
	0.5	0.560	0.558	2.153	2.064	0.052	4.30	0.713	0.709	0.674	0.680
	0.6	0.559	0.561	2.145	2.084	0.058	2.94	0.722	0.717	0.663	0.676
	0.7	0.560	0.559	2.136	2.073	0.055	3.04	0.732	0.724	0.655	0.666
	0.8	0.561	0.562	2.124	2.084	0.059	1.89	0.74	0.734	0.65	0.66
	0.9	0.563	0.563	2.109	2.084	0.055	1.22	0.748	0.745	0.646	0.651
	1.0	0.565	0.567	2.091	2.098	0.061	-0.34	0.756	0.763	0.644	0.639
0.8	0.1	0.747	0.749	3.295	3.225	0.131	2.17	0.790	0.788	0.90	0.910
	0.2	0.722	0.723	3.262	3.114	0.121	4.73	0.791	0.786	0.847	0.858
	0.3	0.709	0.713	3.250	3.099	0.133	4.87	0.795	0.792	0.816	0.831
	0.4	0.702	0.705	3.242	3.069	0.135	5.64	0.801	0.797	0.797	0.810
	0.5	0.699	0.701	3.230	3.059	0.142	5.58	0.808	0.806	0.784	0.793
	0.6	0.698	0.698	3.210	3.036	0.123	5.74	0.814	0.808	0.776	0.786
	0.7	0.698	0.698	3.182	3.037	0.130	4.75	0.821	0.814	0.770	0.780
	0.8	0.699	0.698	3.144	3.028	0.124	3.85	0.827	0.820	0.767	0.773
	0.9	0.700	0.700	3.099	3.042	0.129	1.88	0.832	0.828	0.764	0.768
	1.0	0.702	0.701	3.045	3.038	0.125	0.23	0.837	0.842	0.763	0.756
0.9	0.1	0.873	0.873	6.082	5.876	0.690	3.50	0.895	0.893	0.952	0.957
	0.2	0.859	0.860	6.187	5.826	0.730	6.20	0.895	0.892	0.926	0.934
	0.3	0.852	0.856	6.250	5.918	0.795	5.61	0.897	0.896	0.910	0.920
	0.4	0.848	0.851	6.270	5.909	0.846	6.11	0.900	0.898	0.899	0.910
	0.5	0.846	0.845	6.248	5.756	0.717	8.54	0.904	0.901	0.893	0.897
	0.6	0.846	0.845	6.190	5.777	0.747	7.14	0.907	0.903	0.888	0.894
	0.7	0.845	0.846	6.098	5.804	0.769	5.06	0.910	0.907	0.885	0.891
	0.8	0.846	0.847	5.978	5.861	0.827	2.00	0.913	0.911	0.883	0.889
	0.9	0.846	0.846	5.833	5.785	0.745	0.83	0.916	0.913	0.882	0.885
	1.0	0.847	0.846	5.665	5.735	0.712	-1.24	0.919	0.921	0.881	0.877

by simulation and the relative percentage errors Err(%). The relative errors of the mean number in system are shown to be small and it can be easily seen that $E_{DIS}[L_s] \in [E_{SIM}[L_s] - \sigma_{Ls}, E_{SIM}[L_s] + \sigma_{Ls}]$ for most of the cases and $E_{DIS}[L_s] \in [E_{SIM}[L_s] - 2\sigma_{Ls}, E_{SIM}[L_s] + 2\sigma_{Ls}]$ for the remaining few, where the subscripts "DIS" and "SIM" are clear by context. The maximum relative errors are 1.15%, 2.33%, 4.30%, 5.74%, and 8.74% for ρ=0.5, 0.6, 0.7, 0.8, and 0.9, respectively. Figure 1 shows that the error increases with the traffic intensity and is bigger at middle values (0.5, 0.6) of α for each ρ. The discretization overshoots in almost all cases. The systematic tendency of the relative errors implies, at least partially, that the two residual service times are not independent.



[Figure 1] The relative error of the approximation in terms of E[L_s]

3. Summary

We presented a new intuitive approximation method for finding the steady-state probabilities of the number of customers present in queueing systems with Poisson arrivals and two servers with different deterministic service times and infinite capacity waiting room. We established the system of balance equations based on probabilistic reflections assuming the independence of residual service times of both servers and solved numerically. The balance equations are expressed in terms of the traffic intensity and the ratio of service times of both servers. The method was validated by simulation and turned out to be very sharp.

Acknowledgement

The author would like to thank Professor Henk C. Tijms from the Free University, Amsterdam, the Netherlands, and Professor Abraham Seidmann, Paul Schweitzer, Harry Groenevelt, and Marshall Freimer from the William E. Simon Graduate School of Business Administration, University of Rochester, U.S.A., for their useful comments.

REFERENCES

[1] Ancker, C.J. and A.V. Gafarian, "Queueing with Reneging and Multiple Heterogeneous Servers," *Naval Research Logistics Quarterly*, Vol.10(1963), pp.125-149.

[2] Cooper, R.B., "Queues with Ordered Servers that Work at Different Rates," *Opsearch* Vol.13, No.2(1976), pp.69-78.

[3] Derman, C., G.J. Lieberman, and S.M. Ross, "On the Optimal Assignment of Servers and A Repairman," *Journal of Applied Probability*. Vol.17(1980), pp.577-581.

[4] Gumbel, H., "Waiting lines with heterogeneous server," *Operations Research* Vol.8 (1960), pp.504-511.

[5] Iliadis, I. and L.Y.-C. Lien, "Resequencing Delay Analysis for a Queueing System with Two Heterogeneous servers under a Threshold-Type scheduling," *IEEE Transaction on Communications* Vol.36, No.6

(1988), pp.692-702.

[6] Krishnamoorti, B., "On Poisson Queues with Two Heterogeneous Servers," *Operations Research* Vol.11(1963), pp.341-330.

[7] Larsen, R.L., *Control of Multiple Exponential Servers with Applications to Computer Systems*, Ph. D. Dissertation, Dept. of Computer Science, University of Maryland, 1981.

[8] Larsen, R.L., and A.K. Agrawala, "Control of a Heterogeneous Two-Server Exponential Queueing System," *IEEE Transactions on Software Engineering*, Vol.9, No. 4(1983), pp.522-526.

[9] Lemoine, A., "A Queueing System with Heterogeneous Servers and Autonomous Traffic Control," *Operations Research*, Vol. 23, No.4(1975), pp.681-686.

[10] Lin, W. and P.R. Kumar, "Optimal Control of Queueing System with Two Heterogeneous Servers," *IEEE Transactions on Automatic Control*, Vol.29, No.8(1984), pp. 696-703.

[11] Singh, V.P., "Two-server Markovian queues with balking: heterogeneous vs. homogeneous servers," *Operations Research*, Vol. 18, No.1(1970), pp.145-159.

[12] Singh, V.P., "Markovian queues with three heterogeneous servers," *AIIE Transactions* Vol.3, No.1(1971), pp.45-48.

[13] Tahara, A. and T. Nishida, "Optimal allocation of service rates for multiserver Markovian queue," *Journal of Operations Research Society of Japan*, Vol.18, No.1, 2(1975), pp.90-96.

[14] Tijms, H.C., *Stochastic Modelling and Analysis: A Computational Approach*, Wiley, New York, 1986.

[15] Winston, W., "Optimal Dynamic Rules for Assigning Customers to Servers in A Heterogeneous Queueing System," *Naval Research Logistics Quarterly* Vol.24(1977), pp.293-300.

[16] Winston, W., "Assignment of Customers to Servers in A Heterogeneous Queueing System with Switching," *Operations Research* Vol.25(1977), pp.469-483.

[17] Winston, W., "Optimal Assignment of Customers in a Two-Server Congestion System with No Waiting Room," *Management Science*, Vol.24(1978), pp.702-705.