The Eighth International Symposium on Operations Research and Its Applications (ISORA'09)
Zhangjiajie, China, September 20–22, 2009

A Joining Shortest Queue with MAP Inputs

Zhaotong Lian[1,*] Wenhui Zhou[2] Baohe Su[3]

[1]Faculty of Business Administration, University of Macau, Macau SAR, China
[2]School of Business Administration, South China University of Technology, China
[3]Zhuhai College, Jinan University, Zhuhai, Guangdong, China

Abstract In this paper we study a $MAP/M/K$ system with join-the-shortest-queue (JSQ) discipline which consists of K servers and K queues with infinite buffers. An arrival joins the shortest queue and in the case that there are n ($n \leq K$) queues having the same shortest length, the customer will join any one of these shortest queues with probability $1/n$. We model the system as a level-expanding QBD (LDQBD) process with expanding-block-structured generator matrix by defining the sum of the multi-queue lengths as the level. This method enable us to quantify the regular queueing measures, including stationary joint queue length distribution and sojourn time distribution. We also compare the effect of JSQ and randomly-join-queue.

Keywords Shortest Queue; MAP; Level-expanding; Sojourn Time; Busy Period

1 Introduction

Join-the-shortest-queue (JSQ) policy is an important routing policy within queueing theory. It states that a newly arrived customer will join the shortest queue if there are multiple queues for the same service. The JSQ problem was originally introduced by Haight [6]. The optimality of the JSQ policy has been widely proved. (see [4, 8, 14, 17, 18, 19]). Research on JSQ has attracted many researchers' attention. Knessl et al. [9], Adan et al. [1], Kurkova and Suhov [11], Halfin [7] studied the symmetric JSQ system. Knessl [10], Foley and McDonald [5] studied the asymmetric shortest queue problem. Zhao and Grassmann [20], Adan et al. [2] studied the shortest queue problem with jockeying.

To our knowledge, most study on JSQ system is restricted to Poisson arrivals and exponential service. In this paper we consider the JSQ system with Markovian arrival process (MAP) since MAP is a fairly general process and has a capability of representing a wide class of arrival processes (see [15, 13]). The purpose of this paper is to quantify the regular queueing measures of the $MAP/M/K$ system with JSQ discipline. We model the system as a level-expanding QBD (LDQBD) process with expanding-block-structured generator matrix by defining the sum of the multi-queue lengths as the level. We also develop some approaches and computing algorithms which enable us to analyze performance measures of the system.

This paper is organized as follows. In section 2, we define the JSQ system with MAP inputs. In section 3, we construct an LDQBD process for the joint queue length process

*Corresponding Author. Email: lianzt@umac.mo

and provide a numerical computation method to compute the stationary probability distribution. Section 4 is devoted to the customer sojourn time distribution. We provide some numerical results for some simple examples in section 5.

2 System Definition

We consider a join-the-shortest-queue queueing system with a Markovian arrival process (MAP) with representation (D^0, D^1), where $D^0 = \left(D^0_{i,j}\right)_{m\times m}$ and $D^1 = \left(D^1_{i,j}\right)_{m\times m}$. Let $D = D^0 + D^1$ and z denote its stationary probability vector. Then, $z = (z_1, \cdots, z_m)$ is uniquely determined by $z(D^0 + D^1) = 0$ and $z\mathbf{1} = 1$, and the mean arrival rate of the MAP is $\lambda = zD^1\mathbf{1}$, where $\mathbf{1}$ is a vector with all elements being equal to 1.

The queueing system consists of K servers with unlimited buffers. The service times at all servers are independent and exponentially distributed with rates μ_i for server i, respectively, $i = 1, \cdots, K$. On arrival a customer joins the shortest queue and in the case that there are n queues having the same shortest length, the customer will join any one of these queues with probability $1/n$. Let $\rho = \lambda/(\mu_1 + \cdots + \mu_K)$ denote the traffic intensity of the network. Throughout the paper we assume that the stability condition $\rho < 1$ holds.

Let $N_i(t)$ be the number of customers joining server i (waiting and being served) at time t by the server i, $i = 1, \cdots, K$. Let $\mathbf{N}(t) = (N_1(t), \cdots, N_K(t))$. Denote the phase of the MAP at time t by $S(t)$. Then $\{\mathbf{N}(t), S(t), t \geq 0\}$ is a multi-dimensional continuous-time Markov process with a state space $J = \{(\mathbf{n}, s)$: $\mathbf{n} \in \mathbf{E}^K$, $1 \leq s \leq m\}$, where $\mathbf{E}$ is a set of $\{0, 1, 2, \cdots\}$.

3 Stationary Probability Distribution

Define the system **level** as the total number of customer in the system. We denote it as N, i.e., $N = \mathbf{n1} = \sum_{k=1}^{K} n_k$. We list all possible states of J according to the Level-Ascending and State-Descending (LASD) sequencing rule ([12]). For all level $N \geq 0$, let

$$\begin{aligned} C_{N+1} &= \text{block that the states transit from level } N+1 \text{ to level } N, \\ B_N &= \text{block that the states transit from level } N \text{ to level } N, \text{ and} \\ A_N &= \text{block that the states transit from level } N \text{ to level } N+1. \end{aligned}$$

The corresponding generator matrix Q can be written as

$$Q = \begin{pmatrix} B_0 & A_0 & & & \\ C_1 & B_1 & A_1 & & \\ & C_2 & B_2 & A_2 & \\ & & \ddots & \ddots & \ddots \end{pmatrix}. \tag{1}$$

The blocks in Q are level-dependent, finite, and expanding with the level N.

Let $p = \sum_{k=1}^{K} \chi(n_k = \min\{n_1, \cdots, n_K\})$, where $\chi(\omega)$ is a sign function, i.e., $\chi(\omega) = 0$ if ω is false (or zero); $\chi(\omega) = 1$ if ω is true (or nonzero). Let e_i be a vector of all zero elements except the ith element position which is occupied by 1. Denote $u \in$

$\{v : n_v = \min\{n_1, \cdots, n_K\}\}$. The entries of different blocks in Q are given below (all the remaining elements of Q not mentioned are zeros).

For $\mathbf{n1} = N \geq 0$, and $i, j = 1, \cdots, m$:

$$\begin{aligned}
&\text{in } A_N : (\mathbf{n}, i) \xrightarrow{D^1_{i,j}/p} (\mathbf{n} + e_u, j), \\
&\text{in } C_N : (\mathbf{n}, j) \xrightarrow{\mu_k} (\mathbf{n} - e_k, j), \qquad 1 \leq k \leq K, \\
&\text{in } B_N : (\mathbf{n}, j) \xrightarrow{D^0_{j,j} - \sum_{k=1}^{K} \mu_k \chi(n_k)} (\mathbf{n}, j), \\
&\qquad\qquad (\mathbf{n}, i) \xrightarrow{D^0_{i,j}} (\mathbf{n}, j), \qquad i \neq j.
\end{aligned}$$

With the level dependent generator matrix Q, the following is to calculate the stationary joint distribution $\pi = (\pi(1), \pi(2), \cdots)$, where $\pi(N)$ is the stationary probabilities with level N, $\pi(N) = \{\pi(\mathbf{n}, i) : \mathbf{n1} = N,\ s \in \{1, \cdots, m\}\}$. We adopt an algorithm introduced by Bright and Taylor [3] which is an efficient algorithm to calculate the equilibrium distribution in level dependent quasi-birth-and death processes. We summarize the algorithm briefly. Readers may refer to [3] for details.

Algorithm 1.

Step 1. Truncate Q to Q_{N^*} $(N^* > 0)$ such that $\sum_{N=N^*}^{\infty} \pi(k)e < \varepsilon$.

Step 2. Evaluate the family of matrices $\{R_N,\ N \geq 0\}$ which are the minimal non-negative solutions to the system of equations $A_N + R_N B_{N+1} + R_N[R_{N+1}C_{N+2}] = 0$.

Step 3. Define m_0 to be a positive left eigenvector of the matrix $B_0 + R_0 C_1 = 0$ and define m_N by $m_N = m_0 \Pi_{l=0}^{N-1} R_l$. Let $m = (m_0, m_1, \cdots, m_{N^*})$. Then the equilibrium distribution $\pi = (\pi(1), \pi(2), \cdots, \pi(N^*))$ is given by $\pi(N) = m_N / me$.

The following Theorem presents the value of truncated level N^*.

Theorem 1.

Given an error $\varepsilon > 0$, if $\rho = \frac{\max\{D^1\mathbf{1}\}}{\min\{\mu_1, \cdots, \mu_K\}} < 1$, the largest total queue length N^* can be given by

$$N^* = 1 + \left\lceil \frac{\log(1-\rho) + \log(\varepsilon)}{\log(\rho)} \right\rceil. \tag{2}$$

Proof. Since condition 1 of [3] is held in the JSQ system, we can construct a dominating process of $\{\mathbf{N}(t), S(t), t \geq 0\}$. We define a standard birth-and-death process $\{\bar{I}_N, N \geq 0\}$ on state space $\{N \geq 0\}$ with transition rate $\bar{q}(i, j)$ given by

$$\begin{aligned}
\bar{q}(0,1) &= 0, \\
\bar{q}(N, N+1) &= \max\{A_N\mathbf{1}\} = \max\{D^1\mathbf{1}\},\ \ N > 0, \\
\bar{q}(1,0) &= 0, \\
\bar{q}(N, N-1) &= \min\{C_N\mathbf{1}\} = \min\{\mu_1, \cdots, \mu_K\},\ \ N > 0.
\end{aligned}$$

If $\rho = \frac{\max\{D^1\mathbf{1}\}}{\min\{\mu_1,\cdots,\mu_K\}} < 1$, $\{\bar{l}_N, N \geq 0\}$ is stable. It is sufficient to find N^* such that $\sum_{N=N^*}^{\infty} \bar{l}_N = \rho^{N^*}/(1-\rho) < \varepsilon$, i.e., N^* can be given by (2). According to [3], this N^* is the largest queue length of the JSQ system such that $\sum_{N=N^*}^{\infty} \pi(N)e < \varepsilon$ is satisfied. □

4 The Sojourn Time Distribution

We will derive the distributions of the customer sojourn time in this section. To evaluate the performance of the join-shortest-queue policy, we will compare the customer sojourn time under join-shortest-queue discipline and the randomly-join-queue discipline in which customers randomly join one of the waiting lines without considering the lengths of the waiting line.

4.1 Sojourn Time Distribution under JSQ Discipline

As the initial state of the customer sojourn time is a customer arrival instant, we will first derive the state transition probabilities for the Markov chain embedded at these arrival instants (the epoch right after an arrival).

Because the embedded Markov chain at customer arrival instants is level-dependent $G/M/1$ type, there is not an effective approach to directly derive the corresponding stationary probability distribution. We consider the embedded Markov chain in a system with maximum queue length N^* which can be obtained by Theorem 1.

We decompose the transition generator matrix $\tilde{Q}$ into filtration matrices

$$\tilde{Q}^0 = \begin{pmatrix} B_0 & & & \\ C_1 & B_1 & & \\ & \ddots & \ddots & \\ & & C_{N^*} & \tilde{B}_{N^*} \end{pmatrix}, \tag{3}$$

$$\text{where } \tilde{B}_{N^*}: (\mathbf{n}, j) \xrightarrow{-\sum_{k=1}^{K} \mu_k \chi(n_k)} (\mathbf{n}, j),$$

and

$$\tilde{Q}^1 = \begin{pmatrix} O & A_0 & & & \\ & O & A_1 & & \\ & & \ddots & \ddots & \\ & & & O & A_{N^*-1} \\ & & & & O \end{pmatrix}. \tag{4}$$

The transition probability matrices of the system embedded Markov chain are $\tilde{P}^0 = I - \text{diag}(\tilde{Q})^{-1}\tilde{Q}^0$ and $\tilde{P}^1 = -\text{diag}(\tilde{Q})^{-1}\tilde{Q}^1$ respectively, where $\tilde{P}^1$ represents that a customer arrives, and $\tilde{P}^0$ represents that no customer arrives. Obviously, the transition probability matrix of the embedded Markov chain at a customer arrival instants can be given by

$$X \triangleq \sum_{i=0}^{\infty} [\tilde{P}^0]^i \tilde{P}^1 = -[\tilde{Q}^0]^{-1}\tilde{Q}^1 \tag{5}$$

So X satisfies the equation $\tilde{Q}^0 X = -\tilde{Q}^1$.

To obtain X we use the method of guessing a solution by assuming that

$$X = \begin{pmatrix} O & X_{0,1} & & & \\ O & X_{1,1} & X_{1,2} & & \\ \vdots & \vdots & \vdots & \ddots & \\ O & X_{N^*-1,1} & X_{N^*-1,2} & \cdots & X_{N^*-1,N^*} \\ O & X_{N^*,1} & X_{N^*,2} & \cdots & X_{N^*,N^*} \end{pmatrix} \tag{6}$$

Substituting the guess (6) into $\tilde{Q}^0 X = -\tilde{Q}^1$ yields

$$X_{i,i+1} = -B_i^{-1} A_i \ , \quad i = 0,\ 1,\ 2,\ \cdots, N^* - 1 \tag{7}$$

$$X_{i,j} = (-B_i^{-1} C_i)(-B_{i-1}^{-1} C_{i-1}) \cdots (-B_j^{-1} C_j)(-B_{j-1}^{-1} A_{j-1}) \ , \ \ i = 1, 2, \cdots, N^* - 1,\ j \le i, \tag{8}$$

$$X_{N^*,j} = (-\tilde{B}_{N^*}^{-1} C_{N^*})(-B_{N^*-1}^{-1} C_{N^*-1}) \cdots (-B_j^{-1} C_j)(-B_{j-1}^{-1} A_{j-1}) \ , \ \ j \le N^*, \tag{9}$$

Given an error α, the stationary probability distribution $\tilde{\pi}$ can be easily approximated by $\tilde{\pi} \approx e_1 \tilde{X}^{n^*}$, where n^* is the running times to get the limiting probabilities.

Let $\tilde{\pi}_{\mathbf{n}} = \lim_{t \to +\infty} P\{N(t) = \mathbf{n}\}$ for $\mathbf{n} \in \mathbf{E}^K$. Then the joint probability distribution of the queue length at the epoch of customer arrival, can be given by

$$\tilde{\pi}_{\mathbf{n}} = (\tilde{\pi}_{\mathbf{n},1}, \cdots, \tilde{\pi}_{\mathbf{n},m})\mathbf{1}. \tag{10}$$

Denote the customer sojourn time by W. Let $W_{i,n_i}(x)$ be the probability that $W \le x$ on the condition that there are n_i customers in server i when a customer chooses this server at this customer arrival epoch (including the customer who just arrives to the system). Obviously, $W_{i,n_i}(x)$ is the cumulative probability function of an Erlang distribution with n_i phases.

$$W_{i,n_i}(x) = \int_0^x \frac{\mu_i^{n_i} t^{n_i - 1} exp(-\mu_i t)}{\Gamma(n_i)} dt, \quad x \ge 0. \tag{11}$$

Denote $\underline{i}_{\mathbf{n}} = arg\min\{n_i, i = 1, \cdots, K\}$ and $\underline{n}_{\mathbf{n}} = \min\{n_i, i = 1, \cdots, K\}$. Then the sojourn time distribution under JSQ discipline is

$$P\{W \le x\} = \sum_{N=0}^{\infty} \sum_{\mathbf{n}\mathbf{1}=N} \tilde{\pi}_{\mathbf{n}} W_{\underline{i}_{\mathbf{n}}, \underline{n}_{\mathbf{n}}}(x). \tag{12}$$

4.2 Sojourn Time Distribution under Randomly-join Discipline

In the randomly-join queueing system, customers randomly join one of the waiting lines without considering the lengths of the waiting line and each queue is independent to other queues. If the overall arrival is a Poisson process with arrival rate λ, the arrival process of each queue is a Poisson process with arrival rate λ / K. The lemma below shows that the Markovian arrival process still keeps this property.

Lemma 1.
If the overall customer arrival process is an MAP with parameter (D^0, D^1), and arrival customers randomly join a queue with probability p, then the customer arrival process of this queue is also a MAP with parameter $(D^0 + (1-p)D^1, pD^1)$.

By Lemma 1, if arriving customers randomly join one of queues, each queue will be an $MAP/M/1$ system and the parameter of the MAP is $(\hat{D}^0, \hat{D}^1)$ where $\hat{D}^0 = D^0 + \frac{(K-1)}{K}D^1$ and $\hat{D}^1 = \frac{1}{K}D^1$. The transition matrix generator of the i^{th} queue is $(i = 1, \cdots, K)$

$$Q_R^{(i)} = \begin{pmatrix} \hat{D}^0 & \hat{D}^1 & & & \\ \mu_i I & \hat{D}^0 - \mu_i I & \hat{D}^1 & & \\ & \mu_i I & \hat{D}^0 - \mu_i I & \hat{D}^1 & \\ & & \ddots & \ddots & \ddots \end{pmatrix}. \tag{13}$$

According to Chapter 3 of Neuts [16], it is easy to obtain $\hat{\pi}_i = (\hat{\pi}_{i,1}, \hat{\pi}_{i,2}, \cdots)$, the stationary probability distribution of above $MAP/M/1$ system related to the i^{th} queue. We can also consider it as a special case of Theorem 1 since $Q_R^{(i)}$ is a generator of level-independent QBD:

$$\hat{\pi}_{i,1} = -\hat{\pi}_{i,0}\mu_i^{-1}\hat{D}^0, \tag{14}$$

$$\hat{\pi}_{i,n} = \hat{\pi}_{i,1}\hat{R}_i^{n-1}, \quad n = 2, 3, \cdots \tag{15}$$

where $\hat{R}_i$ satisfies that

$$\hat{D}^1 + \hat{R}_i(\hat{D}^0 - \mu_i I) + \mu_i \hat{R}_i^2 = 0, \tag{16}$$

and $\hat{\pi}_{i,0} = \hat{\tau}_i \hat{z}_i$, in which $\hat{z}_i$ is the stationary probability distribution of $\hat{D}^0 + \mu_i \hat{R}_i$, and $\hat{\tau}_i$ is a scalar uniquely determined by $\sum_{n=0}^{\infty} \hat{\pi}_{i,n}\mathbf{1} = 1$.

Denoted by W_R the sojourn time of a customer who randomly joins one of the queues. We can easily obtain the sojourn time distribution under randomly-join discipline:

$$P\{W_R \le x\} = \frac{1}{K}\sum_{i=1}^{K}\sum_{n=1}^{\infty} \hat{\pi}_{i,n} W_{i,n}(x). \tag{17}$$

5 Numerical Analysis

In this section, we compare the effect of JSQ discipline with randomly-joining queueing discipline through numerical results.

Let $k = 2$. We consider the systems with Poisson arrival process and Markovian arrival process. For the Poisson arrival, we set the arrival rate $\lambda = 0.8$, i.e., $D^0 = -\lambda$ and $D^1 = \lambda$ if we consider it as a special MAP, and the service rates are $\mu_1 = \mu_2 = 2.5$. For the Markovian arrival, we set the parameters $D^0 = \begin{pmatrix} -2 & 0 \\ 0 & -0.5 \end{pmatrix}$ and $D^1 = \begin{pmatrix} 0.6 & 1.4 \\ 0.35 & 0.15 \end{pmatrix}$. The arrival rate is $\lambda = zD^1\mathbf{1} = 0.8$. The service rates are also $\mu_1 = \mu_2 = 2.5$.

In computing the queue length probabilities, we set the accuracy to $\varepsilon = 0.0001$. It is easy to see that the maximum queue buffer is $N^* = 49$ to satisfy $\varepsilon = 0.0001$ according to (2). Fig.1 compares the tail distributions of the steady state sojourn times of Poisson inputs and MAP inputs under JSQ discipline and randomly-join discipline. It shows the JSQ has a significantly lighter tail.

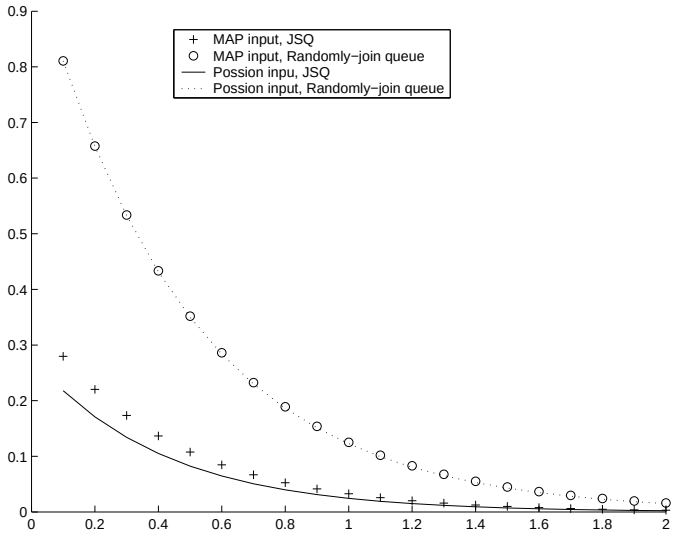

Figure 1: Sojourn time tail distributions ($\lambda = 0.8$, $\mu_1 = \mu_2 = 2.5$)

Acknowledgements

This research is supported in part by the University of Macau through RG003/07-08S/LZT/FBA, the National Natural Science Foundation of China (70871040) and the Natural Science Foundation of Guangdong Province (031924).

References

[1] I. Adan, J. Wessels, and WHM Zijm. Analysis of the symmetric shortest queueing problem. *Stochastic Models*, 6:691–713, 1990.

[2] I. J.-B. F. Adan, J. Wessels, and W. H. M. Zijm. Matrix-geometric analysis of the shortest queue problem with threshold jockeying. *Oper. Res. Lett.*, 13(2):107–112, 1993.

[3] L. Bright and PG Taylor. Calculating the equilibrium distribution in level dependent quasi-birth-and-death processes. *Stochastic Models*, 11(3):497–525, 1995.

[4] A. Ephremides, P. Varaiya, and J. Walrand. A simple dynamic routing problem. *Automatic Control, IEEE Transactions on*, 25(4):690–693, 1980.

[5] RD Foley and DR McDonald. Join the shortest queue: stability and exact asymptotics. *Ann. Appl. Probab*, 11(3):569–607, 2001.

[6] Frank A. Haight. Two queues in parallel. *Biometrika*, 45:401–410, 1958.

[7] S. Halfin. Shortest queue problem. *Journal of Applied Probability*, 22(4):865–878, 1985.

[8] A. Hordijk and G. Koole. On the assignment of customers to parallel queues. *Probability in the Engineering and Informational Sciences*, 6:495–511, 1992.

[9] C. Knessl, B. Matkowsky, Z. Schuss, and C. Tier. Two Parallel Queues with Dynamic Routing. *Communications, IEEE Transactions on [legacy, pre-1988]*, 34(12):1170–1175, 1986.

[10] C. Knessle. A new heavy traffic limit for the asymmetric shortest queue problem. *European Journal of Applied Mathematics*, 10(05):497–509, 1999.

[11] IA Kurkova and Y.M. Suhov. Malyshevą́s theory and JS-queues. Asymptotics of stationary probabilities. *Ann. Appl. Probab*, 13(4):1313–1354, 2003.

[12] Z. Lian and L. Liu. A tandem network with MAP inputs. *Operations Research Letters*, 36(2):189–195, 2008.

[13] D.M. Lucantoni. New results on the single server queue with a batch markovian arrival process. *Stochastic Models*, 7(1):1–46, 1991.

[14] A. Movaghar. Optimal control of parallel queues with impatient customers. *Performance Evaluation*, 60(1-4):327–343, 2005.

[15] M.F. Neuts. A Versatile Markovian Point Process. 1977.

[16] MF Neuts. *Matrix Geometric Solutions in Stochastic Models: An algorithmic Approach.* John Hopkins University Press, 1981.

[17] R. Weber. On the optimal assignment of customers to parallel servers. *Journal of Applied Probability*, 15(2):406–413, 1978.

[18] W. Whitt. Deciding which queue to join: Some counterexamples. *Operations Research*, 34(1):55–62, 1986.

[19] W. Winston. Optimality of the shortest line discipline. *Journal of Applied Probability*, 14(1):181–189, 1977.

[20] Y. Zhao and WK Grassmann. The shortest queue model with jockeying. *Naval research logistics*, 37(5):773–787, 1990.