Proceedings of the 2015 Winter Simulation Conference
L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, eds.

ANALYSIS OF A MAP/PH/1 QUEUE WITH DISCRETIONARY PRIORITY

Ning Zhao

Faculty of Science
Kunming University of Science and Technology,
727 Jingming South Road Kunming,
Yunnan 650500, CHINA

Yaya Guo

Faculty of Science
Kunming University of Science and Technology,
727 Jingming South Road Kunming,
Yunnan 650500, CHINA

Zhaotong Lian

Faculty of Business Administration
University of Macau,Taipa, Macau 999078, CHINA

Mengchang Wang

School of Mechanical and Aerospace Engineering,
Nanyang Technological University, SINGAPORE

ABSTRACT

In this paper, we study a MAP/PH/1 queue with two classes of customers and discretionary priority. There are two stages of service for the low-priority customer. The server adopts the preemptive priority discipline at the first stage and adopts the nonpreemptive priority discipline at the second stage. Such a queueing system can be modelled into a quasi-birth-and-death (QBD) process. But there is no general solution for this QBD process since the generator matrix has a block structure with an infinite number of blocks and each block has infinite dimensions. We present an approach to derive the bound for the high-priority queue length. It guarantees that the probabilities of ignored states are within a given error bound, so that the system can be modelled into a QBD process where the block elements of the generator matrix have finite dimensions. Sojourn time distributions of both high and low priority customers are obtained.

1 INTRODUCTION

Priority mechanism is a scheduling method that allows customers of different classes to receive different quality of service from a single server. It is widely applied in communication systems, health care systems and inventory systems (Brahimi and Worthington 1991, Zhao and Alfa 1995). The priority disciplines in queueing systems can be either preemptive or nonpreemptive. However, both disciplines have some problems. Under the preemptive discipline, a low-priority (LP) job whose service is almost completed may be preempted. And under the nonpreemptive discipline, a high-priority (HP) job may wait even if the LP job with a long service time has just entered service. These unfavorable situations can be avoided by allowing the server to use his discretion to continue or discontinue the service of the LP job (Jaiswal 1968).

Discretionary priority discipline was first introduced by Avi-Itzhak, Brosh, and Naor (1964). They defined a fixed constant z. If the service time of the LP unit is less than z, the preemptive discipline is applied, while if the LP unit has received an amount of service greater than or equal to z, the nonpreemptive discipline is applied. The $M/D/1$ system with two classes of jobs under such discretionary rule was studied and queueing characteristics have been analyzed by using the elementary queueing theory. Melkonian and Kaiser (1996) studied the same discretionary priority queueing discipline for a deterministic service time model and showed that the average sojourn time is shorter than those of preemptive and nonpreemptive models. A number of contributions on the M/G/1 queueing systems with discretionary priorities were made by Jaiswal (1968), Cho and Un (1993), Paterok and Ettl (1994) etc. Kim and Chae (2010) considered a discrete-time discretionary priority queueing model with generally distributed service times. For the

978-1-4673-9743-8/15/$31.00 ©2015 IEEE

preemption mode of the discretionary priority discipline, the authors considered the preemptive resume, preemptive repeat different, and preemptive repeat identical modes. They derived the probability generating functions and the first moments of queue lengths of each class for all the three modes in a unified manner.

All the discretionary priority disciplines defined in the above literature are based on the elapsed or left service time. The server may preempt the current LP customer service and accommodate the HP customer service upon its arrival whenever the preemption condition is satisfied without considering the interrupted cost. In reality, especially in the areas of telecommunication, communication networks and chemical industry, among the entire service procedure, some service stages can be interrupted, while some service stages cannot be interrupted or the interruption cost is very high. In this case, people usually adopt nonpreemptive discipline for the service stage that cannot be interrupted and adopt preemptive discipline for the service stage that can be interrupted (Lian and Zhao 2011, Zhao and Lian 2010, Zhao, Lian, and Wu 2015). In a semiconductor fab, a normal job can be processed when there is no super-hot lot present and is preempted when the super-hot lot arrives under a preemptive policy. However, if the downstream station is busy after the super-hot lot arrives, the super-hot lot can be only processed after the current lot completes its service. Hence, the first station is preemptive and the second station can be non-preemptive. Our model investigates this situation and evaluates its queue time performance.

In this paper we study a MAP/PH/1 queueing system with discretionary priority based on service stages. There are two classes of customers, LP and HP. Both types of customers arrive with Markovian arrival process (MAP) (Neuts 1979, Lucantoni 1991). Each stage of service follows a phase type distribution (PH-distribution) (Neuts 1989). The server adopts the preemptive priority discipline if the LP customer is at the first stage and adopts the nonpreemptive priority discipline if it is at the second stage.

For the priority queueing system, the queue lengths of HP and LP customers can both approach infinity. It is difficult to study this kind of system with the matrix-geometric method directly. Usually people truncate one of the waiting lines into a finite line so that the infinite blocks become finite blocks. In (Alfa 1998, Alfa, Liu, and He 2003), discrete time MAP/PH/1 queues with preemptive and nonpreemptive were studied. The authors truncated those states when the LP queue length becomes sufficiently large and the transition probabilities of those states become sufficiently small, but the error of computational results was difficult to estimate. In this paper we provide another approach to derive the bound of the HP queue length guaranteeing that the probabilities of ignored states are within a given error bound. In this way the matrix-geometric method can be applied.

The paper is organized as follows. In Section 2, we define the MAP/PH/1 queue with discretionary priority based on service stages. In Section 3 we derive the bound of the HP queue length for a given error. The steady-state probability distribution for the system is computed in Section 4. In Section 5, we get the customer sojourn time distribution.We conclude the paper in the last section.

2 THE MAP/PH/1 QUEUE WITH DISCRETIONARY PRIORITY BASED ON SERVICE STAGES

2.1 Arrival process

There are two classes of customers — HP and LP customers. HP customers arrive at the system according to a MAP (Markovian Arrival Process) with representation (D_H^0, D_H^1). D_H^0 and D_H^1 are $m_H \times m_H$ matrices, where m_H denotes the number of states in the underlying Markov Chain which governs HP arrivals. Also, LP customers arrive at the system according to a MAP with representation (D_L^0, D_L^1), where D_L^0 and D_L^1 are $m_L \times m_L$ matrices.

Let $m = m_L m_H$, $D_{01} = D_H^0 \otimes I_{m_L}$, $D_{02} = I_{m_H} \otimes D_L^0$, $D_0 = D_{01} + D_{02}$, $D_1 = D_H^1 \otimes I_{m_L}$ and $D_2 = I_{m_H} \otimes D_L^1$ where $\otimes$ is a Kronecker product (Neuts 1981). It is easy to see that the superposition of the MAPs (D_L^0, D_L^1) and (D_H^0, D_H^1) is a new MAP with representation (D_0, D_1, D_2) ((Li 2009)). D_0 governs the transitions corresponding to no arrivals of both types. D_1 and D_2 govern the transitions corresponding to arrivals of the HP customer and the LP customer respectively.

Let λ_H (resp. λ_L) denote the mean arrival rate of high (resp. low) priority customers. We then have

$$\lambda_H = \pi D_1 \mathbf{e}, \qquad \lambda_L = \pi D_2 \mathbf{e}, \tag{1}$$

where $\mathbf{e}$ denotes a column vector of ones with suitable size and π is the unique $1 \times m$ vector satisfying

$$\pi(D_0 + D_1 + D_2) = 0, \quad \pi\mathbf{e} = 1. \tag{2}$$

2.2 Service process

There is a single server with unlimited waiting buffers. It provides two-stage service for each LP customer. Whenever the first stage service of an LP customer is completed, it directly enters the second stage service. No queue is allowed before the stage-2 service. The service time at stage i follows a phase type distribution (PH-distribution) with representation $(\beta_i,\ S_i)$, $i = 1,2$. Processing time for each HP customer is of phase type represented by $(\beta_3,\ S_3)$. β_i is a vector with size $1 \times l_i$, S_i is a $l_i \times l_i$ matrix, and $S_i^0 = -S_i\mathbf{e}$, $i = 1,2,3$. For simplicity, we call the service of the high priory as stage-3 service. The mean service time at stage i is $\mu_i = -\beta_i S_i^{-1}\mathbf{e},\ i = 1,2,3$.

A discretionary priority service rule is adopted by the server. Whenever an HP unit arrives and finds an LP unit in service, the server takes the following action: if the LP unit is at the first stage service, the preemptive discipline is applied; if the LP unit is in the second stage service, the nonpreemptive discipline is adopted. Among each class, the FCFS rule is adopted. Whenever a preempted LP job is resumed, the service time at stage-1 is a phase type with representation (β_1^*, S_1), which is the stationary distribution of the remaining service time at stage-1, and $\beta_1^*(S_1 + S_1^0\beta_1) = 0,\ \beta_1^*\mathbf{e} = 1$ (see (Neuts 1981)). The vector β_1^* is the limiting probability vector of the phase from which the LP customer starts after it resumes the service. The mean of the remaining service time at stage-1 is $\mu_1^* = -\beta_1^* S_1^{-1}\mathbf{e}$. If $\mu_1 \geq \mu_1^*$, the MAP/PH/1 queue with discretionary priority is stable only if $\lambda_H\mu_3 + \lambda_L(\mu_1 + \mu_2) < 1$.

2.3 System presentation

For simplicity, we denote the MAP/PH/1 system with discretionary priority by $\mathbb{F}$. Let $N_L(t)$ (resp. $N_H(t)$) be the number of the LP (resp. HP) customers in the system at time t, including the customer being serviced by the server. Denote by $\Theta(t)$ the stage of the service at t. Denote by $J(t)$ the phase of the superposition MAP (D_0, D_1, D_2) at time t and $K(t)$ the phase of the service. Then $\mathbb{F}$ can be characterized by a multi-dimensional continuous-time Markov process $Z(t) = \{N_L(t), N_H(t), \Theta(t), J(t), K(t), t \geq 0\}$. The state space $\mathbb{Z} = \Delta_0 \cup \Delta_1 \cup \Delta_2 \cup \Delta_3$, where

$$\begin{aligned}
&\Delta_0 = \{(0,0,0,j,0),\ j = 1,\cdots,m\},\\
&\Delta_1 = \{(0,n_H,3,j,k), n_H > 0;\ j = 1,\cdots,m;\ k = 1,\cdots,l_3\},\\
&\Delta_2 = \{(n_L,0,\theta,j,k), n_L > 0;\ \theta = 1,2;\ j = 1,\cdots,m;\ k = 1,\cdots,l_\theta\},\\
&\Delta_3 = \{(n_L,n_H,\theta,j,k):\ n_L > 0,\ n_H > 0;\ \theta = 2,3;\ j = 1,\cdots,m;\ k = 1,\cdots,l_\theta\}.
\end{aligned}$$

In states $(0,0,0,j,0) \in \Delta_0$, the server is idle with the arrival at the jth phase. In states $(0,n_H,3,j,k) \in \Delta_1$, there is no LP customer and n_H (> 0) HP customers in the system; arrival is in the jth phase and service is in phase k for the HP customer being processed at stage-3. In states $(n_L,0,\theta,j,k) \in \Delta_2$, there are only n_L (> 0) LP customers in the system; the server would serve at stage-1 or 2, and arrival is in the jth phase and service is in phase k with $k = 1,\cdots,l_\theta$, $\theta = 1,2$. In states $(n_L,n_H,\theta,j,k) \in \Delta_3$, both HP and LP customers are present in the system; the arrival process is in the jth phase and the processing is in the kth phase. Note that the HP customer would preempt the stage-1 service of the LP customer, θ could only be 2 and 3 in Δ_3.

Because both N_H and N_L can approach infinity, the block matrices of the generator matrix of the process $Z(t)$ are of infinite dimensions. As mentioned by Alfa (1998), it is difficult to handle a generator with infinite blocks. To solve this problem, we provide an approach to derive the bound of the HP queue length guaranteeing that the probabilities of ignored states are within a given error bound. Sequentially, the stationary probability distribution of the system can be derived by solving a problem with finite-block generator matrix.

3 BOUND OF THE HP QUEUE LENGTH

In this section, we study the bound of queue length for the HP customers. Consider a simpler queueing system denoted by $\hat{\mathbb{F}}$. Assume that $\hat{\mathbb{F}}$ is the same as $\mathbb{F}$ (see Section 2.3) except that there is always one and only one LP customer getting its stage-2 service after an HP customer becomes the leading of its class. Therefore, each HP customer will not start the service until that LP customer finishes the stage-2 service in $\hat{\mathbb{F}}$.

The stationary probability distribution of the HP queue in the system $\hat{\mathbb{F}}$ can be derived without considering the queue of the LP customers. The action of the HP customer in $\hat{\mathbb{F}}$ can be described as a continuous-time Markov process $\hat{Z}(t) = \{\hat{N}_H(t),\, \Theta(t),\, J(t),\, K(t),\, t \geq 0\}$, where $\hat{N}_H(t)$ is the number of the HP customers in the system at time t, including the customer being serviced by the server, $\Theta(t)$ is the stage of the service $(\Theta(t) = 0,2,3)$, $J(t)$ is the phase of (D_{01}, D_1) and $K(t)$ is the phase of the PH-distribution at time t.

We arrange the states in the lexicographic order. The corresponding generator matrix $\widehat{Q}$ can be written as:

$$\widehat{Q} = \begin{pmatrix} \widehat{B}_1 & \widehat{B}_0 & & & \\ \widehat{B}_2 & \widehat{A}_1 & \widehat{A}_0 & & \\ & \widehat{A}_2 & \widehat{A}_1 & \widehat{A}_0 & \\ & & \ddots & \ddots & \ddots \end{pmatrix}, \tag{1}$$

where

$$\widehat{B}_0 = (D_1 \otimes \beta_2 \;\; 0), \quad \widehat{B}_1 = D_{01}, \quad \widehat{B}_2 = (0 \;\; I_m \otimes S_3^0)^T, \tag{2}$$

$$\widehat{A}_0 = \begin{pmatrix} D_1 \otimes I_{l_2} & \\ & D_1 \otimes I_{l_3} \end{pmatrix}, \quad \widehat{A}_2 = \begin{pmatrix} 0 & 0 \\ I_m \otimes (S_3^0 \beta_2) & 0 \end{pmatrix}, \tag{3}$$

$$\widehat{A}_1 = \begin{pmatrix} D_{01} \oplus S_2 & I_m \otimes (S_2^0 \beta_3) \\ 0 & D_{01} \oplus S_3 \end{pmatrix}. \tag{4}$$

$\widehat{Q}$ is tri-diagonal with finite size entries. According to Neuts (1981), the stationary probability distribution of the modified system $\widehat{\pi} = (\widehat{\pi}_0, \widehat{\pi}_1, \cdots)$ is given by

$$\widehat{\pi}_n = \widehat{\pi}_1 \widehat{R}^{n-1}, \quad n = 2,3,\cdots, \tag{5}$$

where $\widehat{R}$ satisfies that

$$\widehat{A}_0 + \widehat{R}\widehat{A}_1 + \widehat{R}^2 \widehat{A}_2 = 0, \tag{6}$$

and the boundary states $\widehat{\pi}_0$ and $\widehat{\pi}_1$ can be determined by solving

$$\widehat{\pi}_0 \widehat{B}_1 + \widehat{\pi}_1 \widehat{B}_2 = 0, \tag{7}$$

$$\widehat{\pi}_0 \widehat{B}_0 + \widehat{\pi}_1 \widehat{A}_1 + \widehat{\pi}_2 \widehat{A}_2 = 0, \tag{8}$$

$$\widehat{\pi}_0 \mathbf{e} + \widehat{\pi}_1 (I - \widehat{R})^{-1} \mathbf{e} = 1. \tag{9}$$

The probability with less than N HP customers in the system is given by

$$\sum_{i=0}^{N} \hat{\pi}_i \mathbf{e} = \hat{\pi}_0 \mathbf{e} + \sum_{i=1}^{N} \widehat{\pi}_1 \widehat{R}^{i-1} \mathbf{e}. \tag{10}$$

As N increases, $\sum_{i=0}^{N} \hat{\pi}_i \mathbf{e}$ approaches 1. For any $\varepsilon > 0$, we can always find a number N_H^* such that $\sum_{i=0}^{N_H^*} \hat{\pi}_i \mathbf{e} > 1 - \varepsilon$.

Denote by N_H and $\hat{N}_H$ the queue length of the HP customers in the system $\mathbb{F}$ and $\hat{\mathbb{F}}$ respectively. Because the arrival processes of HP customers are the same in $\mathbb{F}$ and $\hat{\mathbb{F}}$, and the response time of the HP customers in $\hat{\mathbb{F}}$ is no less than that in $\mathbb{F}$, we have $\hat{N}_H \geq N_H$. For any fixed ε, if there is a number N_H^* such that $P(\hat{N}_H > N_H^*) < \varepsilon$, then $P(N_H > N_H^*) < \varepsilon$. Therefore, we can consider N_H^* as the bound of the HP queue length.

4 STATIONARY PROBABILITY DISTRIBUTION

In the following, we compute the joint steady state distribution of MAP/PH/1 discretionary priority queueing system with maximum HP queue length N_H^*.

Denote by $\mathbb{Z}_k$ the subset of $\mathbb{Z}$, where the queue length of the LP customer is k. If the states in $\mathbb{Z}$ are arranged in the lexicographic order, the generator matrix Q assumes the form

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

The matrices B_i $(i = 0, 1)$ represent the transition from $\mathbb{Z}_0$ to $\mathbb{Z}_{1-i}$, B_2 represents the transition from $\mathbb{Z}_1$ to $\mathbb{Z}_0$. And the matrices A_i $(i = 0, 1, 2)$ represent the transition from $\mathbb{Z}_k$ to $\mathbb{Z}_{k-i+1}$ for $k - i + 1 \geq 1$.

Based on the arrival and service processes described in Section 2, we can figure out the block matrices of the generator matrix Q. Since A_0 corresponds to the arrival of an LP customer, it follows that

$$A_0 = diag(A_0^{00}, A_0^1, A_0^1, \cdots, A_0^1)_{(N_H^*+1)\times(N_H^*+1)}, \tag{2}$$

where

$$A_0^{00} = \begin{pmatrix} D_2 \otimes I_{l_1} & \\ & D_2 \otimes I_{l_2} \end{pmatrix}, \quad A_0^1 = \begin{pmatrix} D_2 \otimes I_{l_2} & \\ & D_2 \otimes I_{l_3} \end{pmatrix}. \tag{3}$$

A_2 corresponds to the departure of an LP customer followed immediately by the start of stage-1 service if no HP customer presents, otherwise stage-3 service starts. The number of HP customers is still the same.

$$A_2 = diag(A_2^{00}, A_2^1, A_2^1, \cdots, A_2^1)_{(N_H^*+1)\times(N_H^*+1)}, \tag{4}$$

where

$$A_2^{00} = \begin{pmatrix} 0 & 0 \\ I_m \otimes (S_2^0 \beta_1) & 0 \end{pmatrix}, \quad A_2^1 = \begin{pmatrix} 0 & I_m \otimes (S_2^0 \beta_3) \\ 0 & 0 \end{pmatrix}. \tag{5}$$

A_1 describes all transitions in which the level remains unchanged. This includes the arrival of an HP customer, the departure of an HP customer followed by the entry into the stage-1 or stage-3 service, changes in the phase of MAP without an LP arrival, and changes in the phase of PH-distribution without an LP departure .

$$A_1 = \begin{pmatrix} A_1^{00} & A_1^{01} & & & & \\ A_1^{10} & A_1^1 & A_1^0 & & & \\ & A_1^2 & A_1^1 & A_1^0 & & \\ & & \ddots & \ddots & \ddots & \\ & & & A_1^2 & A_1^1 & A_1^0 \\ & & & & A_1^2 & A_1^1 + A_1^0 \end{pmatrix}_{(N_H^*+1)\times(N_H^*+1)}, \tag{6}$$

where

$$A_1^{00} = \begin{pmatrix} D_0 \oplus S_1 & I_m \otimes (S_1^0 \beta_2) \\ 0 & D_0 \oplus S_2 \end{pmatrix}, \quad A_1^1 = \begin{pmatrix} D_0 \oplus S_2 & 0 \\ 0 & D_0 \oplus S_3 \end{pmatrix}, \tag{7}$$

$$A_1^{01} = \begin{pmatrix} 0 & D_1 \otimes (\mathbf{e}\beta_3) \\ D_1 \otimes I_{l_2} & 0 \end{pmatrix}, \quad A_1^0 = \begin{pmatrix} D_1 \otimes I_{l_2} & \\ & D_1 \otimes I_{l_3} \end{pmatrix}, \tag{8}$$

$$A_1^{10} = \begin{pmatrix} 0 & 0 \\ I_m \otimes (S_3^0 \beta_1^*) & 0 \end{pmatrix}, \quad A_1^2 = \begin{pmatrix} 0 & 0 \\ 0 & I_m \otimes (S_3^0 \beta_3) \end{pmatrix}. \tag{9}$$

B_0 describes arrivals of LP customers when the system is at level 0.

$$B_0 = diag(B_0^{00}, B_0^1, B_0^1, \cdots, B_0^1)_{(N_H^*+1)\times(N_H^*+1)}. \tag{10}$$

B_0^{00} corresponds to the transitions from $(0,0,0,j,0)$ to $(1,0,\theta,j',k)$, $\theta = 1,\ 2$. B_0^1 corresponds to the transitions from $(0,n_H,3,j,k)$ to $(1,n_H,\theta,j',k')$, $n_H > 0$, $\theta = 2,\ 3$.

$$B_0^{00} = (D_2 \otimes \beta_1 \quad 0), \quad B_0^1 = (0 \quad D_2 \otimes I_{l_3}). \tag{11}$$

B_2 describes the departure of the last LP customer, the server shifts to HP customer if an HP customer is present.

$$B_2 = diag(B_2^{00}, B_2^1, B_2^1, \cdots, B_2^1)_{(N_H^*+1)\times(N_H^*+1)}, \tag{12}$$

where B_2^{00} corresponds to the transitions from $(1,0,\theta,j,k)$ to $(0,0,0,j',0)$, $\theta = 1,\ 2$, B_2^1 corresponds to the transitions from $(1,n_H,\theta,j,k)$ to $(0,n_H,3,j',k')$, $n_H > 0$, $\theta = 2,3$.

$$B_2^{00} = \begin{pmatrix} 0 \\ I_m \otimes S_2^0 \end{pmatrix}, \quad B_2^1 = \begin{pmatrix} I_m \otimes (S_2^0 \beta_3) \\ 0 \end{pmatrix}. \tag{13}$$

Lastly, B_1 describes all the transitions in which the level remains at level 0. If an HP customer arrives at the system, the number of HP customers increases by one. If an HP customer completes its service, and there is at least one HP customer in the waiting line, the server will start the stage-3 service.

$$B_1 = \begin{pmatrix} B_1^{00} & B_1^{01} & & & & \\ B_1^{10} & B_1^1 & B_1^0 & & & \\ & B_1^2 & B_1^1 & B_1^0 & & \\ & & \ddots & \ddots & \ddots & \\ & & & B_1^2 & B_1^1 & B_1^0 \\ & & & & B_1^2 & B_1^1 + B_1^0 \end{pmatrix}_{(N_H^*+1)\times(N_H^*+1)}, \tag{14}$$

where

$$B_1^{00} = D_0, \quad B_1^{01} = D_1 \otimes \beta_3, \quad B_1^{10} = I_m \otimes S_3^0, \tag{15}$$

$$B_1^0 = D_1 \otimes I_{l_3}, \quad B_1^1 = D_0 \oplus S_3, \quad B_1^2 = I_m \otimes S_3^0 \beta_3. \tag{16}$$

With finite-size blocks in Q, we can compute the stationary distribution by the matrix-geometric method as in Section 3. Let $\mathbf{x}$ be the unique solution to $\mathbf{x}Q = 0$ and $\mathbf{xe} = 1$. Partitioning $\mathbf{x}$ as $[\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \cdots]$, where

$\mathbf{x}_i$ is a row vector corresponding to the set of states with i LP customers. $\mathbf{x}_i$ has a matrix-geometric structure as follows:

$$\mathbf{x}_n = \mathbf{x}_1 R^{n-1}, \quad n = 2, 3, \cdots, \tag{17}$$

where R satisfies that

$$A_0 + RA_1 + R^2 A_2 = 0, \tag{18}$$

and the boundary states $\mathbf{x}_0$ and $\mathbf{x}_1$ can be determined by solving

$$\mathbf{x}_0 B_1 + \mathbf{x}_1 B_2 = 0, \tag{19}$$

$$\mathbf{x}_0 B_0 + \mathbf{x}_1 A_1 + \mathbf{x}_2 A_2 = 0, \tag{20}$$

$$\mathbf{x}_0 \mathbf{e} + \mathbf{x}_1 (I - R)^{-1} \mathbf{e} = 1. \tag{21}$$

Define $\mathbf{x}_{n_L}(n_H) = P\{n_H$ HP and n_L LP customers in the system$\}$. $\mathbf{x}_{n_L}(n_H)$ can be expanded at the lower layer as:

$$\mathbf{x}_0(0) = \mathbf{x}_0(0,0), \tag{22}$$

$$\mathbf{x}_0(n_H) = \mathbf{x}_0(n_H, 3), \quad 1 \le n_H \le N_H^*, \tag{23}$$

$$\mathbf{x}_{n_L}(0) = [\mathbf{x}_{n_L}(0, \theta_1), \mathbf{x}_{n_L}(0, \theta_2)], \quad n_L \ge 1, \tag{24}$$

$$\mathbf{x}_{n_L}(n_H) = [\mathbf{x}_{n_L}(n_H, \theta_2), \mathbf{x}_{n_L}(n_H, \theta_3)], \quad 1 \le n_H \le N_H^*, \ n_L \ge 1, \tag{25}$$

where $\mathbf{x}_{n_L}(n_H, \theta_i)$ is a row vector corresponding to the set of states with n_H HP customers, n_L LP customers in the system and stage-i service ($i = 0, 1, 2, 3$).

5 SOJOURN TIME DISTRIBUTIONS

In Section 4, we have derived the stationary distribution of MAP/PH/1 queue with discretionary priority. Now we are ready to determine the sojourn time distribution for LP and HP customers.

5.1 The sojourn time of the HP customer

We start with analyzing the sojourn time distribution of the HP job. An arrival HP customer may find that the system is empty, an LP customer is served at stage 1 or stage 2, or an HP customer is served at stage 3. The sojourn time of an HP customer has nothing to do with the LP customer in the waiting line and the HP customer arrival behind.

Consider the quasi-death process process $Z_H(t) = \{N_H(t), \Theta(t), J(t), K(t), t \ge 0\}$ on state space $\Delta_0' \cup \Delta_1'$, where

$$\Delta_0' = \{(0,0,j,0), \ j = 1, \cdots, m\},$$
$$\Delta_1' = \{(n_H, \theta, j, k): \ n_H > 0; \ \theta = 2,3; \ j = 1, \cdots, m; \ k = 1, \cdots, l_\theta\}.$$

The set Δ_0' consists of the absorbing states where no HP customers are present. Arrange the states of $Z_H(t)$ by its level $N_H(t)$. The infinitesimal generator of $Z_H(t)$ on the state space Δ_1' is

$$T_H = \begin{pmatrix} \widetilde{A}_1 & & & \\ \widetilde{A}_2 & \widetilde{A}_1 & & \\ & \widetilde{A}_2 & \widetilde{A}_1 & \\ & & \ddots & \ddots \end{pmatrix}, \tag{1}$$

where the first block row in T_H corresponds to Δ_1' with $n_H = 1$, the second row to Δ_1' with $n_H = 2$ and so on.

$$\widetilde{A}_1 = \begin{pmatrix} I_m \otimes S_2 & I_m \otimes (S_2^0 \beta_3) \\ 0 & I_m \otimes S_3 \end{pmatrix}, \quad \widetilde{A}_2 = \begin{pmatrix} 0 & 0 \\ 0 & I_m \otimes (S_3^0 \beta_3) \end{pmatrix}. \tag{2}$$

Let W_H be the sojourn time of an arbitrary HP customer and $\bar{w}_{(n,\theta,j,k)}(x) = P\{W_H > x | N_H(0) = n, \Theta(0) = \theta, J(0) = j, K(0) = k\}$ be the conditional tail distribution of the sojourn time of an arrival HP customer who observes the system is in state (n,θ,j,k). Let $\bar{\mathscr{W}}_n(x) = \bar{w}_{(n,:,:,:)}(x)$, $n \geq 1$. Note that the conditional sojourn time of a customer who arrives with the state (n,θ,j,k) is equivalent to the first passage time from state (n,θ,j,k) to an implicit absorbing state $(0,:,:,:)$.

By standard probabilistic arguments, the vector sequence $\{\bar{\mathscr{W}}_n(x)\}$ satisfies the following Kolmogorov's backward equations:

$$\frac{d}{dx}\bar{\mathscr{W}}_1(x) = \widetilde{A}_1 \bar{\mathscr{W}}_1(x), \tag{3}$$

$$\frac{d}{dx}\bar{\mathscr{W}}_n(x) = \widetilde{A}_2 \bar{\mathscr{W}}_{n-1}(x) + \widetilde{A}_1 \bar{\mathscr{W}}_n(x), \quad n \geq 2. \tag{4}$$

We now define $\bar{\mathscr{W}}(x)$ as

$$\bar{\mathscr{W}}(x) = \begin{pmatrix} \bar{\mathscr{W}}_1(x) \\ \bar{\mathscr{W}}_2(x) \\ \vdots \end{pmatrix}. \tag{5}$$

(3)-(4) are then rewritten to be

$$\frac{d}{dx}\bar{\mathscr{W}}(x) = T_H \bar{\mathscr{W}}(x). \tag{6}$$

The solution of (6) is

$$\bar{\mathscr{W}}(x) = \exp(T_H x)\mathbf{e}. \tag{7}$$

We define $\bar{W}(x) = P\{W_H > x\}$ for $x \geq 0$ as the probability that the sojourn time of an arbitrary HP customer in the steady state is greater than x. In order to determine $\bar{W}(x)$, we need to know the probability distribution of the initial state at the arrival epoch of an arbitrary HP customer. Denote by $\alpha(n,\theta,j,k)$ the probability that the system is in state (n,θ,j,k) at the epoch right after a customer arrives. Since the underlying HP arrival process is the MAP with arrival rate λ_H when an arrival takes place, we have

$$\alpha(1,2,:,:) = \lambda_H^{-1} \sum_{n_L=1}^{\infty} \mathbf{x}_{n_L}(0,2)(D_1 \otimes I_{l_2}), \tag{8}$$

$$\alpha(1,3,:,:) = \lambda_H^{-1}\mathbf{x}_0(0,0)(D_1 \otimes \beta_3) + \lambda_H^{-1} \sum_{n_L=1}^{\infty} \mathbf{x}_{n_L}(0,1)(D_1 \otimes (\mathbf{e}\beta_3)), \tag{9}$$

$$\alpha(n_H,2,:,:) = \lambda_H^{-1} \sum_{n_L=1}^{\infty} \mathbf{x}_{n_L}(n_H - 1,2)(D_1 \otimes I_{l_2}), \quad n_H > 1, \tag{10}$$

$$\alpha(n_H,3,:,:) = \lambda_H^{-1} \sum_{n_L=0}^{\infty} \mathbf{x}_{n_L}(n_H - 1,3)(D_1 \otimes I_{l_3}), \quad n_H > 1. \tag{11}$$

The tail distribution of the HP customer sojourn time can be given by

$$\bar{W}(x) = \alpha \bar{\mathscr{W}}(x) = \alpha \exp(T_H x)\mathbf{e}. \tag{12}$$

Applying the uniformization technique (see, for example, (Tijms 1994)), we have, from (7) and (12)

$$\bar{W}(x) = \alpha \sum_{N=0}^{\infty} \frac{h^N x^N}{N!} e^{-hx} \left[I + \frac{T_H}{h} \right]^N \mathbf{e}, \tag{13}$$

where $h = \max\{-(\widetilde{A}_1)_{ii}\}$, $\forall i \geq 1$.

We can calculate $\bar{W}(x)$ by the Rectangle Iterative Algorithm (RIA) (see (Shi, Guo, and Liu 1996)). For $N > 0$, define

$$\widetilde{\alpha}_N = (\alpha(1,:,:,:), \cdots, \alpha(N,:,:,:)), \tag{14}$$

$$\widetilde{T}_{N,N+1} = \begin{pmatrix} I+\frac{1}{h}\widetilde{A}_1 & & & & \mathbf{0} \\ \frac{1}{h}\widetilde{A}_2 & I+\frac{1}{h}\widetilde{A}_1 & & & \mathbf{0} \\ & \ddots & \ddots & & \\ & & \frac{1}{h}\widetilde{A}_2 & I+\frac{1}{h}\widetilde{A}_1 & \mathbf{0} \end{pmatrix}_{N\times(N+1)}, \tag{15}$$

$$\widetilde{T}^{(N)}_{N_0} = \widetilde{T}_{N_0,N_0+1}\widetilde{T}_{N_0+1,N_0+2}\cdots\widetilde{T}_{N_0+N-1,N_0+N}. \tag{16}$$

Algorithm The RIA for $\bar{W}(x)$.

Step 1. Given an error $\delta > 0$ and a positive number Y, let $\varepsilon = \delta/[2+hY]$;

Step 2. Find N_0, so that $\sum_{N=1}^{N_0} \alpha(N,:,:,:)\mathbf{e} > 1-\varepsilon$;

Step 3. Let

$$L = \min\left\{\max([2heY],[log_2(1/\varepsilon)]), \inf\{N|\widetilde{\alpha}_{N_0}\widetilde{T}^{(N)}_{N_0}\mathbf{e} \leq \varepsilon\}\right\};$$

Step 4. $\forall x \in [0,Y]$, calculate

$$\bar{W}_{N_0,L}(x) = \sum_{n=0}^{L} \frac{h^n x^n}{n!} e^{-hx} \widetilde{\alpha}_{N_0} \widetilde{T}^{(n)}_{N_0} \mathbf{e}. \tag{17}$$

According to Shi et al. (1996), $\bar{W}_{N_0,L}(x)$ approximates $\bar{W}(x)$ for $x \in [0,Y]$ with a uniform error δ.

5.2 The sojourn time of the LP customer

In this section, we specify the relevant matrices to obtain an algorithm of the sojourn time distribution of the LP customer. The determination procedures are the same as Section 5.1.

When an LP customer arrives, it may find that the system is empty, an LP customer is served at the stage 1 or stage 2, or an HP customer is served at stage 3. Therefore, this customer needs to wait whenever the server is not idle. Note that the HP customers arriving behind may affect the sojourn time of this LP customer because of the preemption rule. In order to analyze the sojourn time of an LP arrival customer, we need to consider the HP arrivals, the number of customers in the waiting line and the stage of the present service.

We construct another quasi-death process $Z_L(t) = \{N_L(t), N_H(t), \Theta(t), J(t),\ K(t), t \geq 0\}$ on the state space $\Delta_0'' \cup \Delta_1'' \cup \Delta_2''$, where

$$\begin{aligned} \Delta_0'' &= \{(0,0,0,j,0),\ j = 1,\cdots,m\}, \\ \Delta_1'' &= \{(n_L,0,\theta,j,k), n_L > 0;\ \theta = 1,2;\ j = 1,\cdots,m;\ k = 1,\cdots,l_\theta\}, \\ \Delta_2'' &= \{(n_L,n_H,\theta,j,k):\ n_L > 0,\ 0 < n_H \leq N_H^*;\ \theta = 2,3;\ j = 1,\cdots,m; \\ &\qquad k = 1,\cdots,l_\theta\}. \end{aligned}$$

The state set Δ_0'' consists of the absorbing states where no low priority customers are present. Arrange the states of $Z_L(t)$ by its level $N_L(t)$. We denote by A_1' the block of states that transit from level i to level i. For $i \geq 2$, we denote by A_2 the block of states that transit from level i to level $i-1$. We obtain the generator of the process $\{Z_L(t),\ t > 0\}$ on the states space $\Delta_1'' \cup \Delta_2''$:

$$T_L = \begin{pmatrix} A_1' & & & \\ A_2 & A_1' & & \\ & A_2 & A_1' & \\ & & \ddots & \ddots \end{pmatrix}, \tag{18}$$

where A_1' equals A_1 replacing D_0 by D_{01}. A_1 and A_2 have been defined in Section 4.

Denote by $\alpha'(n_L, n_H, \theta, j, k)$, the probability that the system is in state (n_L, n_H, θ, j, k) at the epoch right after an LP customer arrives.

$$\alpha'(1,0,1,:,:) = \lambda_L^{-1}\mathbf{x}_0(0,0)(D_2 \otimes \beta_1), \tag{19}$$

$$\alpha'(1,n_H,2,:,:) = \mathbf{0}, \quad 0 \leq n_H \leq N_H^*, \tag{20}$$

$$\alpha'(1,n_H,3,:,:) = \lambda_L^{-1}\mathbf{x}_0(n_H,3)(D_2 \otimes I_{l_3}),\ 1 \leq n_H \leq N_H^*, \tag{21}$$

$$\alpha'(n_L,0,:,:,:) = \left[\lambda_L^{-1}\mathbf{x}_{n_L-1}(0,1)(D_2 \otimes I_{l_1}),\ \ \lambda_L^{-1}\mathbf{x}_{n_L-1}(0,2)(D_2 \otimes I_{l_2})\right], \quad n_L > 1, \tag{22}$$

$$\alpha'(n_L,n_H,:,:,:) = \left[\lambda_L^{-1}\mathbf{x}_{n_L-1}(n_H,2)(D_2 \otimes I_{l_2}),\ \ \lambda_L^{-1}\mathbf{x}_{n_L-1}(n_H,3)(D_2 \otimes I_{l_3})\right], \quad n_L > 1,\ 1 \leq n_H \leq N_H^*. \tag{23}$$

The tail sojourn time distribution of the LP customer is given by

$$\bar{W}'(x) = P\{W_L > x\} = \alpha' \exp(T_L x)\mathbf{e}. \tag{24}$$

By using the Rectangle Iterative Algorithm in Section 5.1, we can compute the tail distribution of sojourn time of an arbitrary LP customer.

6 CONCLUSION

In this paper, a service system with discretionary priority based on service stages is studied. Given an error bound, the system can be approximated by a system with finite HP queue length, so that the stationary probability distribution of the system can be derived by the matrix-geometric method. The sojourn time distribution of an arbitrary customer is obtained.

The proposed model can be used to estimate the job queue time for systems with discretionary priority based on service stages in semiconductor fabs if the job arrival processes are MAPs and service times follow PH-distribution. In addition to our simple model, the practical situation can be more complicated: a station may suffer different types of interruptions (Wu, McGinnis, and Zwart 2008, Wu, McGinnis, and Zwart 2011a, Wu 2014a), and face parallel or serial batches (Wu 2014b, Wu, McGinnis, and Zwart 2011b, Wu, Wu, Zhao, and Xu 2014) with different impact on queue time and productivity (Wu, McGinnis, and Zwart 2007). Under this situation, service time is not identical to process time (Wu and Hui 2008) and system variability increases (Wu 2005). Together with our model which considers job priority, all above factors will affect the performance of the station and its downstream stations (Wu and McGinnis 2012, Wu and Zhao 2015), which can be captured by the intrinsic ratio (Wu and McGinnis 2013). The impact of queue time under all factors is left for future research.

REFERENCES

Alfa, A. 1998. “Matrix-Geometric Solution of Discrete Time MAP/PH/1 Priority Queue”. *Naval Research Logistics* 45 (1): 23–50.

Alfa, A., B. Liu, and Q. He. 2003. “Discrete-Time Analysis of MAP/PH/1 Multiclass General Preemptive Priority Queue”. *Naval Research Logistics* 50 (6): 662–682.

Avi-Itzhak, B., I. Brosh, and P. Naor. 1964. “On Discretionary Priority Queueing”. *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik* 44 (6): 235–242.

Brahimi, M., and D. Worthington. 1991. “Queueing Models for Out-Patient Appointment Systems —— A Case Study”. *The Journal of the Operational Research Society* 42 (9): 733–746.

Cho, Y., and C. Un. 1993. “Analysis of the M/G/1 Queue under a Combined Preemptive/Nonpreemptive Priority Discipline”. *IEEE Transactions on Communications* 41 (1): 132–141.

Jaiswal, N. 1968. *Priority Queues*. Academic Press.

Kim, K., and K. Chae. 2010. “Discrete-Time Queues with Discretionary Priorities”. *European Journal of Operational Research* 200 (2): 473–485.

Li, Q. 2009. *Constructive Computation in Stochastic Models with Applications*. Berlin: Springer.

Lian, Z., and N. Zhao. 2011. “A Two-Stage M/G/1 Queue with Discretionary Priority”. In *International Conference on Industrial Engineering and Engineering Management*, 1402–1406. IEEE.

Lucantoni, D. 1991. “New Results on the Single Server Queue with a Batch Markovian Arrival Process”. *Stochastic Models* 7 (1): 1–46.

Melkonian, V., and M. Kaiser. 1996. “Discretionary Priority Discipline: A Reasonable Compromise Between Preemptive and Nonpreemptive Disciplines”. *Applied Mathematics Letters* 9 (4): 91–94.

Neuts, M. 1979. “A Versatile Markovian Point Process”. *Journal of Applied Probability* 16 (4): 764–779.

Neuts, M. 1981. *Matrix Geometric Solutions in Stochastic Models: An algorithmic Approach*. John Hopkins University Press.

Neuts, M. 1989. *Structured Stochastic Matrices of M/G/1 Type and Their Applications*. Marcel Decker Inc., New York.

Paterok, M., and M. Ettl. 1994. “Sojourn Time and Waiting Time Distributions for M/GI/1 Queues with Preemption-Distance Priorities”. *Operations Research* 42 (6): 1146–1161.

Shi, D., J. Guo, and L. Liu. 1996. “SPH-Distributions and the Rectangle-Iterative Algorithm”. *Matrix-Analytic Methods in Stochastic Models*:207–224.

Tijms, H. 1994. *Stochastic Models: An Algorithmic Approach*. Chichester: Wiley.

Wu, K. 2005. “An examination of variability and its basic properties for a factory”. *IEEE Transactions on Semiconductor Manufacturing* 18 (1): 214–221.

Wu, K. 2014a. “Classification of Queueing Models for a Workstation with Interruptions: A Review”. *International Journal of Production Research* 52 (3): 902–917.

Wu, K. 2014b. “Taxonomy of Batch Queueing Models in Manufacturing Systems”. *European Journal of Operational Research* 237 (1): 129–135.

Wu, K., and K. Hui. 2008. “The determination and indetermination of service times in manufacturing systems”. *IEEE Transactions on Semiconductor Manufacturing* 21 (1): 72–82.

Wu, K., and L. McGinnis. 2012. “Performance evaluation for general queueing networks in manufacturing systems: Characterizing the trade-off between queue time and utilization”. *European Journal of Operational Research* 221 (2): 328–339.

Wu, K., and L. McGinnis. 2013. “Interpolation approximations for queues in series”. *IIE transactions* 45 (3): 273–290.

Wu, K., L. McGinnis, and B. Zwart. 2011a. “Queueing Models for a Single Machine Subject to Multiple Types of Interruptions”. *IIE transactions* 43 (10): 753–759.

Wu, K., L. F. McGinnis, and B. Zwart. 2007. “Compatibility of Queueing Theory, Manufacturing Systems and SEMI Standards”. In *IEEE International Conference on Automation Science and Engineering*, 501–506. IEEE.

Wu, K., L. F. McGinnis, and B. Zwart. 2008. “Queueing models for single machine manufacturing systems with interruptions”. In *Winter Simulation Conference Miami, FL, USA*, 2083–2092. IEEE.

Wu, K., L. F. McGinnis, and B. Zwart. 2011b. “Approximating the Performance of a Batch Service Queue Using the $M/M^k/1$ Model”. *Automation Science and Engineering, IEEE Transactions on* 8 (1): 95–102.

Wu, K., Z. Wu, N. Zhao, and Y. Xu. 2014. “Approximating the performance of a station subject to changeover setups”. In *Proceedings of the 2014 Winter Simulation Conference*, 2396–2403. IEEE Press.

Wu, K., and N. Zhao. 2015. “Dependence among single stations in series and its applications in productivity improvement”. *European Journal of Operational Research* DOI:10.1016/j.ejor.2015.05.028.

Zhao, N., and Z. Lian. 2010. “A Two-Stage Discretionary Priority Service System with Markovian Arrival Inputs”. In *International Conference on Industrial Engineering and Engineering Management*, 438–442. IEEE.

Zhao, N., Z. Lian, and K. Wu. 2015. “Analysis of a MAP/PH/1 Queue with Discretionary Priority Based on Service Stages”. *Asia-Pacic Journal of Operational Research* DOI:10.1142/S0217595915500426.

Zhao, Y., and A. Alfa. 1995. “Performance Analysis of a Telephone System with Both Patient and Impatient Customers”. *Telecommunication Systems* 4 (1): 201–215.

AUTHOR BIOGRAPHIES

NING ZHAO is an assistant professor in the Faculty of Science at Kunming University of Science and Technology. She received her Ph.D. degree in Data-Base Management and Information System from University of Macau in 2011. Her current research interests are queueing systems and inventory control. Her email address is zhaoning@kmust.edu.cn.

YAYA GUO is a graduate student in the Faculty of Science at Kunming University of Science and Technology. Her e-mail address is 742149130@qq.com.

ZHAOTONG LIAN is a professor in the Faculty of Business Administration, University of Macau. His research interest is queueing systems. He has published many research papers in outstanding journals, such as Operations Research, Mathematics of Operations Research. His email address is lianzt@umac.mo.

MENGCHANG WANG is a postdoctoral research fellow at Nanyang Technological University. He holds Ph.D. degree in Industrial Engineering from Huazhong University of Science and Technology. His email address is wangmc@ntu.edu.sg.