Stability Conditions of a Preemptive Repeat Priority Queue with Customer Transfers

Qi-Ming He[1], Jingui Xie[2], Xiaobo Zhao[2]

[1] Department of Industrial Engineering, Dalhousie University
Halifax, N.S., Canada B3J 2X4
Email: qi-ming.he@dal.ca

[2] Department of Industrial Engineering
Tsinghua University, Beijing, China
Emails: xiejingui@tsinghua.org.cn and xbzhao@tsinghua.edu.cn

Abstract: This paper is concerned with the stability of a preemptive repeat priority queueing system with multiple types of customers and customer transfers between queues. Using the mean-drift method and matrix-analytic methods, simple conditions for the queueing system to be stable or instable are found. The conditions indicate that system stability depends only on the service rate of customers of the highest priority and arrival rates of individual types of customers. That implies that the service rates and the transfer rates of all lower priority customers have no impact on system stability - a result that can be useful in the design of such queueing systems.

Keywords: Priority queue; Markov chain; Mean-drift method; Stability; Ergodicity; Matrix-analytic methods.

1 Introduction

Queueing models have found extensive applications in manufacturing, telecommunications, and service industries. Queueing models also find applications in healthcare, public safety, and social justice systems (Larson, 1987). In a hospital emergency department, patients are categorized into critical and non-critical groups. A patient in the critical group will be attended by a doctor, if one is available, as soon as the patient arrives. The condition of a patient in the non-critical group may deteriorate while waiting, and become critical. Then the patient has to be attended as soon as a doctor is available. In a fire/911 department, emergence cases are also categorized. The dispatch of ambulances and fire-trucks is arranged accordingly. In the design of such systems, allocation of resource is a key issue, especially the allocation of some scarce resource. One of the goals of resource allocation is to ensure that the system is stable in the sense that the queue lengths will not grow too long. Motivated by the applications in healthcare and public safety sectors, this paper introduces a preemptive repeat priority $MMAP[N]/PH[N]/S$ queue with customer transfers and finds its stability/instability conditions.

Queueing systems with customer priorities and queueing systems with customer transfers have been studied extensively. Existing works address issues related to system stability, optimal scheduling, routing, and performance analysis (e.g., Adan, Wessels, and Zijm, 1991, Whitt, 1986, and Zhao and Grassmann, 1995). Some of the existing works focus on system stability conditions, some on the stationary analysis of the queue length(s) and waiting times, and some on

customer transfer strategies.

In the preemptive repeat priority *MMAP*[*N*]/*PH*[*N*]/*S* queue of interest, only lower priority customers can transfer to higher priority customers. The model is different from those in the existing literature. Consequently, the stability/instability conditions of the system are different from that of the existing models. The results obtained in this paper imply that system stability/instability depends only on the service rate of customers of the highest priority and arrival rates of individual types of customers. That implies that system stability/instability is independent of the service rates and transfer rates of lower priority customers. The results also imply that the correlations between the arrival processes of customers have no impact on system stability/instability. The results can be useful in the design of such queueing systems. In Xie, He, and Zhao (2008), similar results are shown for a simpler model *M*[*N*]/*M*[*N*]/*S*, where customers arrive according to independent Poisson processes and service rates and transfer rates are exponentially distributed. Compared to that of Xie, He, and Zhao (2008), the queueing model considered in this paper has fairly general assumptions on its arrival process, service times, and transfer times. Thus, this model captures features, such as the correlations between arrivals, that do not exist in the simpler model. Therefore, the applicability of the results is extended significantly.

The remainder of the paper is organized as follows. The queueing model of interest is introduced in Section 2. In Section 3, the main results – stability and instability conditions – are presented. A brief discussion on the proof of the main results is given as well.

2 Queueing model

The queueing model of interest consists of S identical servers serving N types of customers: type 1, type 2, …, and type N customers. Type 1, 2, …, and N customers form queue 1, 2, …, and N, respectively. Type N customers have the highest service priority, type N–1 the second highest service priority, …, and type 1 the lowest service priority. The priority level of a type k customer is k. The S servers are numbered as server 1, 2, …, and S.

A type k customer can transfer to a type k+1 customer while it is waiting for service, for $1 \le k \le N-1$. The time for a waiting type k customer to be transferred into a type k+1 customer is called the transfer time. The clock of the transfer time of a customer is set to zero and begins to click as soon as the customer joins a queue waiting for service. A customer in service does not change its type or its priority level.

Customers are served on a preemptive repeat basis. That implies that an interrupted service is repeated. We also assume that the clock of the transfer time of a customer whose service is interrupted is reset to zero. The service discipline is specified as follows.

a) Suppose that, when a type k customer arrives, some of the servers are idle. Then the type k customer enters one of the idle servers and begins its service immediately. Exactly which server to enter does not affect system stability analysis.

b) Suppose that, when a type k customer arrives, all servers are busy. If the priority level of all customers in service is k or higher, then the type k customer joins queue k. The clock of the customer's transfer time is set to zero and begins to click. If the priority level of some customers in service is lower than k, then one of the customers of the lowest priority in service is pushed out of its server and back into its queue, and the server begins to serve the type k customer immediately. The clock of the transfer time of the customer pushed out is reset to zero and begins to click. The service of this customer will be repeated when the customer enters a server later. Exactly which customer of the lowest priority in service is pushed out does not affect system stability analysis, since the service times of lower priority customers do not affect system stability (see Theorem 1).

c) Suppose that, when a server completes a service, at least one queue is not empty. The server chooses a customer from the nonempty queue of the highest priority and begins to serve it immediately. Exactly which customer to be chosen from that queue does not affect system stability analysis. For mathematical convenience, a method for the server to choose a customer will be specified, after the distribution of the transfer time is defined later in this section.

d) Suppose that, when a server completes a service, all queues are empty. Then the server becomes idle.

e) Suppose that, when a type k customer transfers to a type k+1 customer, there are type k customers in service. Then one of the type k customers in service is pushed back into queue k. The server begins to serve the transferred customer immediately. The clock of the transfer time of the type k customer just pushed out is reset to zero and begins to click. That customer will repeat it service when it enters a server later.

Next, we define the arrival process, service times, and transfer times explicitly.

The arrival process The N types of customers arrive according to a marked Markov arrival process (*MMAP*[N]) (see Neuts, 1979, He and Neuts, 1998). The *MMAP*[N] has a matrix representation $\{D_0, D_J, J\in\Phi\}$, where Φ is a set of strings of integers defined as

$$\Phi = \{J: \quad J = j_1 j_2 \cdots j_N, \text{ where } j_1, j_2, ..., j_N \geq 0, J \neq 0, D_J \neq 0\}, \qquad (1)$$

D_0 and $\{D_J, J\in\Phi\}$ are matrices of order m_a, D_0 is a matrix with negative diagonal elements and nonnegative off-diagonal elements, $\{D_J, J\in\Phi\}$ are nonnegative elements, $\left(D_0 + \sum_{J\in\Phi} D_J\right)\mathbf{e} = 0$, and $\mathbf{e}$ is a column vector with all elements being one. The matrix D_J, $J\in\Phi$, is for the arrival rates of type J batches that include j_1 type 1 customers, j_2 type 2 customers, …, and j_N type N customers, conditioning on the phase of a underlying continuous time Markov chain (CTMC) just prior to the arrival. Let $D = D_0 + \sum_{J\in\Phi} D_J$. Then D is the infinitesimal generator of the underlying CTMC of the arrival process. We assume that the matrix D is irreducible, i.e., the underlying CTMC is irreducible. Let $I_a(t)$ be the phase of the underlying CTMC at time t. Denote by $\boldsymbol{\theta}_a$ the nonnegative row vector satisfying

$\boldsymbol{\theta}_a D = 0$ and $\boldsymbol{\theta}_a \mathbf{e} = 1$. Since D is irreducible, every element of $\boldsymbol{\theta}_a$ is positive. Then the stationary arrival rate of type k customers is given by $\lambda_k = \boldsymbol{\theta}_a \sum_{J \in \Phi} j_k D_J \mathbf{e}$, for $1 \le k \le N$.

Define $D^*(z) = D_0 + \sum_{J \in \Phi} z^{|J|} D_J$, where $|J| = j_1 + j_2 + \ldots + j_N$, which is the number of customers (regardless of their types) in the batch J. We assume that there exists $\hat{z} > 1$ such that $D^*(z)$ is a finite matrix for $0 < z < \hat{z}$. This assumption is not restrictive, since it is satisfied if the set Φ has a finite number of elements or the batch size has a discrete phase-type distribution (Neuts, 1981).

To make it easy to understand *MMAP*[*N*], we give two examples of *MMAP*[*N*].

Example 2.1 Assume that all customers arrive individually, i.e., all batch sizes are one. For this case, Φ = {10···0, 010···0, …, 0···01}. A string J = 0···010···0, whose k-th number is 1, represents a batch that has a single type k customer in it.

Example 2.2 Assume that N=2 and Φ = {10, 01, 11, 22}. For this case, customers arrive in four forms: a single type 1 arrival (J=10) , a single type 2 arrival (J=01), a batch with one type 1 customer and one type 2 customer (J=11), and a batch with 2 type 1 customers and 2 type 2 customers (J=22).

MMAP[*N*] is a versatile process that can be used to model complicated multi-type arrival processes with correlations between individual arrivals and/or with special arrival patterns. According to Asmussen and Koole (1993), *MMAP*[*N*] can approximate any multi-type arrival processes.

The service times The service times of the type k customers have the same phase-type distribution with a *PH*-representation $(\boldsymbol{\alpha}_k, T_k)$ of order m_k, $1 \le k \le N$, where $\boldsymbol{\alpha}_k$ is a stochastic vector, i.e., $\boldsymbol{\alpha}_k$ is nonnegative and $\boldsymbol{\alpha}_k \mathbf{e} = 1$ (which implies that the service time is positive with probability 1), and T_k is a *PH*-generator, i.e., T_k is invertible, diagonal elements of T_k are negative, off-diagonal elements of T_k are nonnegative, and the vector $\mathbf{T}_k^0 = -T_k \mathbf{e}$ is nonnegative. The mean service time of type k customers is $\mu_k^{-1} = -\boldsymbol{\alpha}_k T^{-1} \mathbf{e}$, $1 \le k \le N$. Then μ_k is the service rate of type k customers. Without loss of generality, we assume that the *PH*-representation $(\boldsymbol{\alpha}_k, T_k)$ is irreducible, which is equivalent to that the infinitesimal generator $T_k + \mathbf{T}_k^0 \boldsymbol{\alpha}_k$, is irreducible. The irreducibility of the *PH*-representation is assumed to ensure that a CTMC to be defined for the queue length processes is irreducible. Let $\boldsymbol{\theta}_k$ be the nonnegative vector satisfying $\boldsymbol{\theta}_k (T_k + \mathbf{T}_k^0 \boldsymbol{\alpha}_k) = 0$ and $\boldsymbol{\theta}_k \mathbf{e} = 1$. Then all elements of $\boldsymbol{\theta}_k$ are positive. In fact, it can be verified that $\boldsymbol{\theta}_k = -\mu_k \boldsymbol{\alpha}_k T_k^{-1}$. Let $I_i(t)$ be the phase of the underlying CTMC of the service undergoing in server i at time t, $1 \le i \le S$. If server i is idle, we define $I_i(t) = 0$. Note that the range of $I_i(t)$ depends on the type of the customer in service at time t. We refer to Neuts (1981) for more about phase-type distributions.

The transfer times The transfer time of a type k customer to a type k+1 customer has a Coxian distribution with a Coxian representation $(\boldsymbol{\beta}_k, S_k)$ of order n_k, $1 \le k \le N-1$, where $\boldsymbol{\beta}_k = (\beta_{k,1}, \beta_{k,2}, \ldots, \beta_{k,n_k})$ is a stochastic vector (i.e., $\boldsymbol{\beta}_k \ge 0$ and $\boldsymbol{\beta}_k \mathbf{e} = 1$, which implies that the transfer time is positive with probability one) and S_k is a Coxian generator given as follows:

$$S_k = \begin{pmatrix} -s_{k,1} & s_{k,1} & & & & \\ & -s_{k,2} & s_{k,2} & & & \\ & & \ddots & \ddots & & \\ & & & -s_{k,n_k-1} & s_{k,n_k-1} & \\ & & & & -s_{k,n_k} \end{pmatrix}. \quad (2)$$

It is easy to see that Coxian distribution is a special case of *PH*-distribution. Without loss of generality, we assume that $\beta_{k,1}>0$ for $1\le k\le N-1$. Similar to the irreducibility condition made on the *PH*-representations of service times, this assumption is made to ensure that a CTMC to be defined for the queue length processes is irreducible. For the Coxian distribution, the phase of its underlying CTMC always moves up by one at a transition epoch. This property is used in the definitions of the queue length processes. For type N customers, there is no transfer. We introduce a underlying CTMC with a single phase and no transition, i.e., $n_N = 1$, $s_{N,1} = 0$, and $\beta_{N,1}=1$.

We assume that the arrival process, service times, and transfer times are independent.

3. Main results

Based on the above definitions, we define a CTMC for queue 1, queue 2, …, and queue N. For any type k customers waiting for service, the underlying CTMC of its transfer time must be in a state i, $1\le i \le n_k$. Thus, the queue k of type k customers can be decomposed into $n_k + 1$ subqueues: queue $(k, 0)$ consists of type k customers in service, queue $(k, 1)$ consists of type k customers waiting in queue k and the underlying CTMCs of their transfer times are in state 1, …, and queue (k, n_k) consists of type k customers waiting in queue k and the underlying CTMCs of their transfer times are in state n_k. When a server is available to serve a type k customer, we assume that one of the type k customers whose underlying CTMC is in the highest phase is chosen for service.

Define $q_{k,i}(t)$ the number of type k customers in queue (k, i) at time t, $0\le i\le n_k$, $k = 1, 2, \ldots, N$. Let $\mathbf{q}_k(t) = (q_{k,0}(t), q_{k,1}(t), \ldots, q_{k,n_k}(t))$, $1\le k\le N$. Because of the preemption property, we must have $q_{j,i}(t) = 0$ for $j \ge k+1$ and $1 \le i \le n_j$, if $q_{k,0}(t) > 0$. Define

$$\mathbf{q}(t) = (\mathbf{q}_1(t), \mathbf{q}_2(t), \ldots, \mathbf{q}_{N-1}(t), \mathbf{q}_N(t)), \text{ and}$$
$$\mathbf{X}(t) = (\mathbf{q}(t), I_a(t), I_1(t), \ldots, I_s(t)).$$

The first part of $\mathbf{X}(t)$ (i.e., $\mathbf{q}(t)$) provides information on the lengths of the N queues as well as transferring times of customers in queues. The rest of $\mathbf{X}(t)$ (i.e., $(I_a(t), I_1(t), \ldots, I_s(t))$) provides information on the underlying phases of the arrival process and service times. It can be verified, under our assumptions, the CTMC $\{\mathbf{X}(t), t \ge 0\}$ is irreducible. Denote by Ω the state space of $\{\mathbf{X}(t), t \ge 0\}$. A typical state in Ω has the form $\mathbf{x} = (\mathbf{q}, i_a, i_1, \ldots, i_s)$ with $\mathbf{q} = (\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_{N-1}, \mathbf{q}_N)$ and $\mathbf{q}_k = (q_{k,0}, q_{k,1}, \ldots, q_{k,n_k})$, $1\le k\le N$.

We call the queueing system *stable* if the CTMC $\{\mathbf{X}(t), t \ge 0\}$ is ergodic (irreducible and positive recurrent). The CTMC is called non-ergodic if it is not

ergodic. We call the queueing system *instable* if the CTMC $\{\mathbf{X}(t), t \geq 0\}$ is non-ergodic. The ergodicity of the CTMC $\{\mathbf{X}(t), t \geq 0\}$ is characterized in the following theorem.

Theorem 1 Assume that D is irreducible, $D^*(z)$ is a finite matrix for $0 < z < \hat{z}$ with $\hat{z} > 1$, the *PH*-representations of all service times are *PH*-irreducible, and $\beta_{k,1} > 0$ for $1 \leq k \leq N$. Then the CTMC $\{\mathbf{X}(t), t \geq 0\}$ is irreducible.

1.1) The CTMC $\{\mathbf{X}(t), t \geq 0\}$ is ergodic if $\sum_{k=1}^{N} \lambda_k < S\mu_N$.

1.2) The CTMC $\{\mathbf{X}(t), t \geq 0\}$ is non-ergodic if $\sum_{k=1}^{N} \lambda_k > S\mu_N$.

Part 1.1) and part 1.2) indicate that the ergodicity/non-ergodicity conditions of the CTMC (or the stability/instability of the queueing system) are independent of the service rates and the transfer rates of lower priority customers. In addition, the correlations between individual arrival processes, the service rates of lower priority customers, and the transfer rates of lower priority customers have no impact on system stability.

A complete proof of Theorem 1 can be found in He, Xie, and Zhao (2009). In the proof, matrix-analytic methods (Neuts 1981) and the mean-drift method (Meyn and Tweedie, 1996) are utilized. Theorem 1.18 for ergodicity of Markov chains given in Chen (1991) and Theorem 1 for non-ergodicity given in Choi and Kim (2004) are applied.

References

Asmussen, S. and G. Koole. 1993. Marked point processes as limits of Markovian arrival streams, *J. Appl. Prob*. 30: 365-372.

Chen, M. F. 1991. On Three Classical Problems for Markov Chains with Continuous Time Parameters. *Journal of Applied Probability* 28(2): 305-320.

Choi, B.D. and B. Kim 2004. Non-ergodicity criteria for denumerable continuous time Markov processes. *Operations Research Letters* 32: 574-580.

He, Q.M. and M.F. Neuts 1998. Markov chains with marked transitions, *Stochastic Processes and their Applications* 74(1): 37-52.

He, Q.M., J.G. Xie, and X.B. Zhao 2009. Stability Conditions of a preemptive repeat priority *MMAP*[*N*]/*PH*[*N*]/*S* queue with customer transfers. Working paper #09-01, 2009, Department of Industrial Engineering, Dalhousie University.

Larson, R. C. 1987. Perspectives on queues: social justice and the psychology of queueing, *Operations Research* 35(6): 895-905.

Meyn, S.P. and R. Tweedie 1993. *Markov Chains and Stochastic Stability*, Springer Verlag.

Neuts, M.F. 1979. A versatile Markovian point process, *Journal of Applied Probability* 16: 764-779.

Neuts, M.F. 1981. *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach*, The Johns Hopkins University Press, Baltimore.

Whitt, W. 1986. Deciding which queue to join: some counterexamples, *Operations Research* 34: 55-62.

Xie, J.G., Q.M. He, and X.B. Zhao 2008. Stability of a priority queueing system with customer transfers, *OR Letters* Vol 36: 705-709.

Zhao, Y. and W.K. Grassmann 1995. Queueing analysis of a jockeying model, *Operations Research* 43: 520-529.