Ergodicity of the BMAP/PH/s/$s+K$ retrial queue with PH-retrial times

Qi-Ming He [a], Hui Li [b] and Yiqiang Q. Zhao [c]

[a] *Department of Industrial Engineering, Dalhousie University, Halifax, NS, Canada B3J 2X4*
E-mail: qi-ming.he@dal.ca

[b] *Department of Mathematics, Mount Saint Vincent University, Halifax, NS, Canada B3M 2J6*
E-mail: hli@msvu1.msvu.ca

[c] *Department of Mathematics and Statistics, University of Winnipeg, Winnipeg, MB, Canada R3B 2E9*
E-mail: zhao@UWinnipeg.ca

Received 24 March 1999; revised 21 November 1999

Define the traffic intensity as the ratio of the arrival rate to the service rate. This paper shows that the BMAP/PH/s/$s+K$ retrial queue with PH-retrial times is ergodic if and only if its traffic intensity is less than one. The result implies that the BMAP/PH/s/$s+K$ retrial queue with PH-retrial times and the corresponding BMAP/PH/s queue have the same condition for ergodicity, a fact which has been believed for a long time without rigorous proof. This paper also shows that the same condition is necessary and sufficient for two modified retrial queueing systems to be ergodic. In addition, conditions for ergodicity of two BMAP/PH/s/$s+K$ retrial queues with PH-retrial times and impatient customers are obtained.

Keywords: retrial queue, ergodicity, sample path method, mean-drift method, matrix analytic methods, batch Markov arrival process (BMAP), PH-distribution, impatient customer, Lyapunov function

1. Introduction

The primary retrial queueing system studied in this paper has K waiting positions and s servers. When an arriving customer finds that all servers are busy and no waiting position is available, the customer enters an orbit and retries for service after a random time until the customer gets into service or the queue. The orbit can accommodate any number of orbiting customers. We focus on the ergodicity problem of a large class of such retrial queueing systems where the retrial times are non-exponential. We also study the ergodicity problem of several variants of the primary retrial queueing system, such as retrial queueing systems with impatient customers and retrial queueing systems with a finite number of retrial customers.

Ergodicity of retrial queues has been studied by many researchers (see [1,10–12,14,23], and references therein). By that a queueing system is ergodic, we mean that an associated Markov process (defined later) of the queueing system is ergodic. Con-

ditions for ergodicity have been obtained for various retrial queueing systems. Denote by ρ the traffic intensity defined as the ratio of the arrival rate to the (total) service rate of a retrial system. In [10], the sufficiency of $\rho < 1$ for ergodicity of the M/M/s/s retrial queue with exponential retrial times was proved. Liang and Kulkarni [16] obtained a stability condition for a single server retrial queue. Diamond [5] and Yang et al. [22] showed that $\rho < 1$ is a necessary and sufficient condition for ergodicity of the M/G/1 retrial queue with general retrial times, independently. For single server retrial queues with a Markov arrival process, PH-service times, and exponential retrial times (or their special cases), Diamond [5], Diamond and Alfa [6,7], and Li and Yang [15] proved that $\rho < 1$ is necessary and sufficient for ergodicity. Diamond and Alfa [8,9] proved that $\rho < 1$ is a sufficient condition for ergodicity of multiserver retrial queues with finite buffers and exponential retrial times. These results support a simple and intuitive conjecture that a retrial queueing system is ergodic if and only if $\rho < 1$. Unfortunately, the conjecture is not always true. A counterexample is shown below (see [16] for more examples).

Example 1.1. Consider a single server retrial queueing system with deterministic interarrival times and deterministic retrial times, both of which have the same length one. There is no waiting position in this queueing system. The service time is 0.1 with probability 0.9 and 5.1 with probability 0.1. The mean service time is 0.6. Thus, the traffic intensity ρ is 0.6, which is less than one. For this queueing system, assume that an arrival instant is followed by a (possible) retrial. The system is unstable since the number of customers in the orbit increases to infinity. The reason for this is that with a positive probability an arrival will find that the server is busy and therefore the customer has to enter the orbit. On the other hand, no customer in the orbit can enter service. Furthermore, the capacity of the queueing system is wasted since the server is often idle for 0.9 units of time while many customers are in the orbit.

The conjecture does not hold in the above example because the interarrival times and retrial times are discrete. However, it is still believed that for a retrial queue whose interarrival times, service times, and retrial times have continuous distribution functions, the conjecture holds. The objective of this paper is to show that $\rho < 1$ is a necessary and sufficient condition for ergodicity of the BMAP/PH/s/s + K retrial queue with PH-retrial times. The BMAP/PH/s/s + K retrial queue with PH-retrial times has a batch Markov arrival process (BMAP), PH-service times, s ($< \infty$) servers, K ($< \infty$) waiting positions, and PH-retrial times, which can be considered as a generalization of the retrial queues studied in [5–10,15,21]. Among them, the MAP/PH/s/s + K retrial queue with exponential retrial times considered in [8,9] is the one closest to the model studied in this paper.

In this paper, the sample path approach is used to prove the necessity of the condition $\rho < 1$, and the mean-drift method (Falin and Templeton [12] and also see Foster's criterion in [4]) is used to prove the sufficiency of the condition. The sample path approach is different and more rigorous compared to other methods used to prove

the necessity conditions for retrial queues [12]. Although Foster's criterion has been adopted as a standard way to prove the sufficiency of the ergodicity conditions for retrial queues, the extension from the case with exponential retrial times and single arrivals to the case with PH-retrial times and batch arrivals is not trivial and is in fact challenging. It is worth pointing out that the study of retrial queueing systems with non-exponential retrial times has been very limited, and this paper makes a contribution in this area. Besides the main theorem, conditions for two modified retrial queues and two retrial queues with impatient customers to be ergodic are obtained as well. Results obtained in this paper can be used to determine whether or not a retrial queuing system can reach its steady state and to choose system parameters, such as the number of servers, to ensure system stability.

The rest of the paper is organized as follows. In section 2, the BMAP/PH/s/$s + K$ retrial queue with PH-retrial times is defined. In section 3, a Markov process is introduced to represent the queueing system and the main theorem is stated. In sections 4 and 5, condition $\rho < 1$ is proved to be necessary and sufficient for ergodicity of the BMAP/PH/$s/s + K$ retrial queue with PH-retrial times, respectively. In section 6, two retrial queues with impatient customers are defined and conditions for ergodicity are obtained. In section 7, two modified retrial queueing systems of the BMAP/PH/$s/s + K$ retrial queue with PH-retrial times are proved to be ergodic if and only if $\rho < 1$. Finally, some discussion is given to the results obtained in this paper in section 8.

2. The BMAP/PH/$s/s + K$ retrial queue with PH-retrial times

The primary queueing model under consideration in this paper is defined in this section. First, the input process – a batch Markov arrival process – is introduced. Then the service time of a customer is defined and the retrial mechanism is specified.

Customers arrive at the queueing system according to a batch Markov arrival process (BMAP). The BMAP was introduced by Neuts (see [17,18,20]) as a generalization of the phase-type renewal process (see [19]). It is defined on a finite irreducible Markov process $I(t)$ (called the underlying Markov process) which has m states and an infinitesimal generator D. In the BMAP, the sojourn time in state i is exponentially distributed with parameter $(-D_0)_{i,i}$ $(\geqslant -(D)_{i,i})$, $1 \leqslant i \leqslant m$. At the end of the sojourn time in state i, there occurs a transition to another (possibly the same) state and that transition may or may not correspond to the arrival of customers. Let D_k be the rate matrix of transitions that generate an arrival with k customers, $0 \leqslant k \leqslant \infty$. Notice that matrix D_0 has strictly negative diagonal elements and nonnegative off-diagonal elements, matrices $\{D_n,\ n \geqslant 1\}$ are nonnegative, and $D = D_0 + \sum_{n\geqslant 1} D_n$. Let θ be the stationary probability vector of the underlying Markov process $I(t)$, i.e., θ satisfies $\theta D = 0$ and $\theta \mathbf{e} = 1$, where $\mathbf{e}$ is a column vector of ones. The stationary arrival rate is then given by $\lambda = \theta \sum_{n=1}^{\infty} nD_n\mathbf{e}$ (which is assumed to be finite). Define $D^*(z) = \sum_{n=0}^{\infty} z^n D_n$. We assume that $D^*(z)$ is finite for $0 < z < z_0$, where $z_0 > 1$. The assumption holds when a finite number of matrices in $\{D_n,\ n \geqslant 1\}$ are nonzero,

but it becomes restrictive when an infinite number of matrices in $\{D_n,\ n \geqslant 1\}$ are nonzero. For instance, the assumption excludes batch size distributions with a heavy tail.

There are s identical servers and each serves one customer at a time. Service times of customers are independent of each other and have a common phase-type distribution (PH-distribution) function with a matrix representation (α, T), where α is a nonnegative vector of size m_1 with $\alpha\mathbf{e} = 1$, and T is an $m_1 \times m_1$ matrix. Let $\mathbf{T}^0 = -T\mathbf{e}$. The mean service time is given by $1/\mu = -\alpha T^{-1}\mathbf{e}$ and μ is the average service rate of a server. For more details about the PH-distribution, see [19, chapter 2]. When a service is complete, the customer leaves the queueing system immediately and the server becomes available to serve another customer in the queue (if any).

There are K waiting positions, where K is a nonnegative integer. Thus, there are at most $s + K$ customers present in the system at any time. When a customer arrives and finds an idle server, the customer receives service immediately. When a customer arrives and finds that all servers are busy and a waiting position is available, the customer occupies the waiting position. Otherwise, the customer waits for a random period of time for retrial. When a customer is waiting for retrial, the customer is considered to be in an "orbit" and the retrial is independently identically (probabilistically) repeated until a server or a waiting position is seized. The retrial times have a PH-distribution with matrix representation (β, H), $\beta\mathbf{e} = 1$, where β is a nonnegative vector of size m_2, and $H = (h_{i,j})$ is an $m_2 \times m_2$ matrix. Define $\mathbf{H}^0 = -H\mathbf{e}$. Then $\mathbf{H}^0$ is a nonnegative column vector. Denote by h_i^0 the ith element of vector $\mathbf{H}^0$. Notice that when there are a number of customers in the orbit, the next customer entering service or taking a waiting position does not have to be the customer who entered the orbit first. Clearly, any customer in the orbit must be in one of the m_2 states of the PH-distribution at any time.

Throughout this paper, we assume that the service times, retrial times, and the input process are mutually independent.

3. The infinitesimal generator

In this section, a Markov process is constructed to represent the BMAP/PH/s/$s + K$ retrial queue with PH-retrial times. Let

$N_i(t)$ be the number of customers in the orbit whose retrial process is in state i at time t, $1 \leqslant i \leqslant m_2$;

$q(t)$ be the total number of customers in queue or in service at time t;

$I(t)$ be the state of the underlying Markov process of the BMAP at time t;

$I_j(t)$ be the state of the service time of the jth working server at time t, $1 \leqslant j \leqslant \min\{s, q(t)\}$.

It is easy to see that $\{N_i(t),\ 1 \leqslant i \leqslant m_2,\ q(t),\ I(t),\ I_j(t),\ 1 \leqslant j \leqslant \min\{s, q(t)\}\}$ is an irreducible Markov process. Let $\mathcal{N} = \{\mathbf{n} = (n_1, \ldots, n_{m_2})\colon\ 0 \leqslant n_i < \infty,\ 1 \leqslant i \leqslant m_2\}$. The state space of the Markov process is

$$\Omega = \bigcup_{\mathbf{n}\in\mathcal{N}} \Omega_{\mathbf{n}}, \tag{3.1}$$

where $\Omega_{\mathbf{n}} = \Omega_{\mathbf{n},0} \cup \Omega_{\mathbf{n},1} \cup \Omega_{\mathbf{n},2} \cup \cdots \cup \Omega_{\mathbf{n},s+K}$, and $\Omega_{\mathbf{n},q} = \{(\mathbf{n}, q)\} \times \{1, 2, \ldots, m\} \times \{1, 2, \ldots, m_1\}^q$ if $0 \leqslant q < s$; otherwise (i.e., $s \leqslant q \leqslant s + K$), $\Omega_{\mathbf{n},q} = \{(\mathbf{n}, q)\} \times \{1, 2, \ldots, m\} \times \{1, 2, \ldots, m_1\}^s$. The subset $\Omega_{\mathbf{n}}$ is called level $\mathbf{n}$. Each level has $M \equiv m + mm_1 + mm_1^2 + \cdots + mm_1^s + Kmm_1^s$ states, where $\equiv$ means a definition equation. In each state in $\Omega_{\mathbf{n}}$ with $\mathbf{n} = (n_1, \ldots, n_{m_2})$, there are $n_1 + \cdots + n_{m_2}$ customers in the orbit.

For convenience, transitions of the Markov process $\{N_i(t),\ 1 \leqslant i \leqslant m_2,\ q(t),\ I(t),\ I_j(t),\ 1 \leqslant j \leqslant \min\{s, q(t)\}\}$ are described in terms of the transitions between levels. Let $\mathbf{e}_i$ be a vector of size m_2 with all elements zero except that the ith element is one, $1 \leqslant i \leqslant m_2$, and let vector $\mathbf{k} = (k_1, \ldots, k_{m_2})$ with all elements nonnegative integers. From level $\mathbf{n}$, the Markov process can move to level $\mathbf{n} + \mathbf{k}$, $\mathbf{n} - \mathbf{e}_i$, or $\mathbf{n} - \mathbf{e}_i + \mathbf{e}_j$, for $1 \leqslant i,\ j \leqslant m_2$, in one transition. The infinitesimal generator of Markov process $\{N_i(t),\ 1 \leqslant i \leqslant m_2,\ q(t),\ I(t),\ I_j(t),\ 1 \leqslant j \leqslant \min\{s, q(t)\}\}$ is given as follows.

From level $\mathbf{n}$ to level $\mathbf{n} + \mathbf{k}$ ($\mathbf{k} \neq 0$), the matrix of transition rates is given by

$$\begin{aligned} A(\mathbf{n}, \mathbf{n}+\mathbf{k}) = &\begin{pmatrix} 0 & \ldots & 0 & p(\mathbf{k},\beta)D_{s+K+\widehat{K}(\mathbf{k})} \otimes \alpha \otimes \alpha \otimes \cdots \otimes \alpha \\ 0 & \ldots & 0 & p(\mathbf{k},\beta)D_{s-1+K+\widehat{K}(\mathbf{k})} \otimes \mathbf{I}_{m_1} \otimes \alpha \otimes \cdots \otimes \alpha \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \ldots & 0 & p(\mathbf{k},\beta)D_{1+K+\widehat{K}(\mathbf{k})} \otimes \mathbf{I}_{m_1^{s-1}} \otimes \alpha \\ 0 & \ldots & 0 & p(\mathbf{k},\beta)D_{K+\widehat{K}(\mathbf{k})} \otimes \mathbf{I}_{m_1^s} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \ldots & 0 & p(\mathbf{k},\beta)D_{\widehat{K}(\mathbf{k})} \otimes \mathbf{I}_{m_1^s} \end{pmatrix} \\ &\equiv p(\mathbf{k},\beta)A_0\big(\widehat{K}(\mathbf{k})\big), \end{aligned} \tag{3.2}$$

where

$$\widehat{K}(\mathbf{k}) = \sum_{j=1}^{m_2} k_j \quad \text{and} \quad p(\mathbf{k},\beta) = \frac{(k_1 + \cdots + k_{m_2})!}{k_1! \cdots k_{m_2}!}\, \beta_1^{k_1} \cdots \beta_{m_2}^{k_{m_2}}, \tag{3.3}$$

$x! = x(x-1)\cdots 1$, $0! = 1$, $\otimes$ denotes the Kronecker product [13], and $\mathbf{I}$ ($\mathbf{I}_n$) is the $(n \times n)$ identity matrix. $A(\mathbf{n}, \mathbf{n} + \mathbf{k})$ is an $M \times M$ matrix. Note that the blocks within $A(\mathbf{n}, \mathbf{n} + \mathbf{k})$ describe the transition from sublevels $\{\Omega_{\mathbf{n},0}, \Omega_{\mathbf{n},1}, \ldots, \Omega_{\mathbf{n},s+K}\}$ to $\{\Omega_{\mathbf{n}+\mathbf{k},0}, \Omega_{\mathbf{n}+\mathbf{k},1}, \ldots, \Omega_{\mathbf{n}+\mathbf{k},s+K}\}$. When a batch of n customers arrives, some of the

n customers fill idle servers and the queue first, and the rest of them enter the orbit. For those who enter the orbit, the selection of the initial states of their retrial times follows the multinomial distribution $\{p(\mathbf{k}, \beta)\}$.

From level $\mathbf{n}$ to level $\mathbf{n} - \mathbf{e}_i + \mathbf{e}_j$, $1 \leqslant i, j \leqslant m_2$,

$$\begin{aligned} &A(\mathbf{n}, \mathbf{n} - \mathbf{e}_i + \mathbf{e}_j) = n_i h_{i,j} \mathbf{I}, \quad \text{if } 1 \leqslant i \neq j \leqslant m_2,\ n_i > 0; \\ &A(\mathbf{n}, \mathbf{n}) = A_1 + \sum_{i=1}^{m_2} n_i h_{i,i} \mathbf{I} + \sum_{i=1}^{m_2} n_i h_i^0 (\mathbf{I} - \Gamma), \end{aligned} \tag{3.4}$$

where Γ is an $M \times M$ diagonal matrix whose first $M - mm_1^s$ elements have a value one and last mm_1^s elements have a value zero, $n_i h_{i,j} \mathbf{I}$ is the matrix of the total transition rates from state i to j $(i \neq j)$ for the retrial process, $n_i h_i^0 (\mathbf{I} - \Gamma)$ represents the matrix of the total transition rates that retrial customers find a full queue upon finishing the retrial time in state i, and A_1 represents the matrix of the transition rates due to an arrival or service completion:

$$A_1 = \begin{pmatrix} B_{0,0} & B_{0,1} & B_{0,2} & \cdots & B_{0,s+K} \\ B_{1,0} & B_{1,1} & B_{1,2} & \cdots & B_{1,s+K} \\ & \ddots & \ddots & \ddots & \vdots \\ & & B_{s+K-1,s+K-2} & B_{s+K-1,s+K-1} & B_{s+K-1,s+K} \\ & & & B_{s+K,s+K-1} & B_{s+K,s+K} \end{pmatrix}, \tag{3.5}$$

where

$$B_{i,j} = \begin{cases} D_{j-i} \otimes \mathbf{I}_{m_1^i} \otimes \underbrace{\alpha \otimes \cdots \otimes \alpha}_{j-i}, & 0 \leqslant i < j \leqslant s, \\ D_{j-i} \otimes \mathbf{I}_{m_1^i} \otimes \underbrace{\alpha \otimes \cdots \otimes \alpha}_{s-i}, & 0 \leqslant i < s \leqslant j \leqslant s + K, \\ D_{j-i} \otimes \mathbf{I}_{m_1^s}, & s \leqslant i < j \leqslant s + K; \end{cases} \tag{3.6}$$

$$B_{i,i} = \begin{cases} D_0 \otimes \mathbf{I}_{m_1^i} + \sum_{j=0}^{i-1} \mathbf{I}_{mm_1^j} \otimes T \otimes \mathbf{I}_{m_1^{i-1-j}}, & 0 \leqslant i \leqslant s, \\ D_0 \otimes \mathbf{I}_{m_1^s} + \sum_{j=0}^{s-1} \mathbf{I}_{mm_1^j} \otimes T \otimes \mathbf{I}_{m_1^{s-1-j}}, & s+1 \leqslant i \leqslant s+K; \end{cases} \tag{3.7}$$

$$B_{i,i-1} = \begin{cases} \mathbf{I}_m \otimes \left(\sum_{j=0}^{i-1} \mathbf{I}_{m_1^j} \otimes \mathbf{T}^0 \otimes \mathbf{I}_{m_1^{i-1-j}} \right), & 1 \leqslant i \leqslant s, \\ \mathbf{I}_m \otimes \left(\sum_{j=0}^{s-1} \mathbf{I}_{m_1^j} \otimes \left(\mathbf{T}^0 \alpha \right) \otimes \mathbf{I}_{m_1^{s-1-j}} \right), & s+1 \leqslant i \leqslant s+K. \end{cases} \tag{3.8}$$

Note that $[\sum_{n \geqslant 1} A_0(n) + A_1]\mathbf{e} = 0$, and $A(\mathbf{n}, \mathbf{n})$ and $A(\mathbf{n}, \mathbf{n} - \mathbf{e}_i + \mathbf{e}_j)$ are $M \times M$ matrices. Equation (3.6) represents the matrix of the transition rates corresponding to

an arrival. For an arrival of size $j - i$ ($j \leqslant s + K$), $\min\{s - i, j - i\}$ customers enter service and the rest of them join the waiting line. Equation (3.7) represents the matrix of the transition rates without an arrival or a service completion. Equation (3.8) represents the matrix of the transition rates corresponding to a service completion, which describes the change of states when a service is complete and a new service begins.

From level $\mathbf{n}$ to level $\mathbf{n} - \mathbf{e}_i$, for $1 \leqslant i \leqslant m_2$ and n_i of the vector $\mathbf{n}$ is positive,

$$
\begin{aligned}
&A(\mathbf{n}, \mathbf{n} - \mathbf{e}_i) \\
&= n_i h_i^0 \begin{pmatrix}
0 & \mathbf{I}_m \otimes \alpha & & & & & & \\
& 0 & \mathbf{I}_{mm_1} \otimes \alpha & & & & & \\
& & \ddots & \ddots & & & & \\
& & & 0 & \mathbf{I}_{mm_1^{s-1}} \otimes \alpha & & & \\
& & & & 0 & \mathbf{I}_{mm_1^s} & & \\
& & & & & \ddots & \ddots & \\
& & & & & & 0 & \mathbf{I}_{mm_1^s} \\
& & & & & & & 0
\end{pmatrix} \\
&\equiv n_i h_i^0 A_2. \qquad (3.9)
\end{aligned}
$$

Note that $A(\mathbf{n}, \mathbf{n} - \mathbf{e}_i)$ is an $M \times M$ matrix and $(A_2 - \Gamma)\mathbf{e} = 0$. Also notice that a retrial is successful only when the total number of customers in queue or service is less than $s + K$. Although the infinitesimal generator is complicated, its construction is straightforward and explicit.

When the Markov process introduced above is ergodic, we say that the retrial queue is ergodic. The main objective of this paper is to prove the following theorem.

Theorem 1. Let $\rho = \lambda/(s\mu)$ be the traffic intensity of the queueing system. The BMAP/PH/s/$s + K$ retrial queue with PH-retrial times defined in section 2 is ergodic if and only if $\rho < 1$.

The proof of theorem 1 consists of two parts: (1) a proof of the necessity of the condition and (2) a proof of the sufficiency of the condition, which are provided in the following two sections.

4. Proof of the necessity

To prove the necessity of the condition $\rho < 1$ in theorem 1, the sample path method is utilized. The BMAP/PH/s queue considered here is the classical queueing system (with infinite waiting positions and no retrial) which has the same input process,

service times, and the number of servers as in the retrial queueing system defined in section 2.

Theorem 2. If the BMAP/PH/s/$s+K$ retrial queue with PH-retrial times defined in section 2 is ergodic, then the corresponding BMAP/PH/s queue is ergodic. This implies that $\rho < 1$.

Proof. The sample path approach is used to prove this theorem. The idea is to compare the queue length of the retrial queue of interest to that of the BMAP/PH/s queue. Suppose that the retrial queue and the corresponding nonretrial queue BMAP/PH/s are empty initially. Let their queueing processes be coupled in the same probability space. Let a_n be the arrival epoch of the nth customer and s_n the service time of the nth service. Notice that s_n may not be the nth customer's service time in the retrial queue. Let t_n and $t_{L,n}$ be the epochs at which the nth service starts for the retrial queue and the corresponding nonretrial queue, respectively. It is clear that, for $n > s$,

$$\begin{aligned} t_n &\geqslant \max\{a_n, \min\{t_{n-s} + s_{n-s}, t_{n-s+1} + s_{n-s+1}, \ldots, t_{n-1} + s_{n-1}\}\}; \\ t_{L,n} &= \max\{a_n, \min\{t_{L,n-s} + s_{n-s}, t_{L,n-s+1} + s_{n-s+1}, \ldots, t_{L,n-1} + s_{n-1}\}\}. \end{aligned} \tag{4.1}$$

By induction, it is easy to prove that $t_n \geqslant t_{L,n}$ and $t_n + s_n \geqslant t_{L,n} + s_n$ for $n > 0$. Let $A(t)$ be the total number of customers who arrived in $(0,t)$. Let $B(t)$ and $B_L(t)$ be the total number of services finished in $(0,t)$ and let $q_{\text{all}}(t)$ and $q_{L,\text{all}}(t)$ be the total number of customers in service, queue, or the orbit for the retrial and the nonretrial queues, respectively. It is clear that

$$\begin{aligned} &A(t) = \max\{n\colon\ a_n < t\}; \\ &B(t) = \max\{n\colon\ t_n + s_n < t\}, \qquad B_L(t) = \max\{n\colon\ t_{L,n} + s_n < t\}; \\ &q_{\text{all}}(t) = A(t) - B(t), \qquad q_{L,\text{all}}(t) = A(t) - B_L(t). \end{aligned} \tag{4.2}$$

It is easy to see that $B(t) \leqslant B_L(t)$ and, therefore, $q_{\text{all}}(t) \geqslant q_{L,\text{all}}(t)$ for all t. This implies that the total number of customers in the BMAP/PH/s/$s+K$ retrial queue is always as large as that in the BMAP/PH/s queue. This further implies that $\mathbf{P}\{q_{\text{all}}(t) \leqslant q\} \leqslant \mathbf{P}\{q_{L,\text{all}}(t) \leqslant q\}$ for all $q \geqslant 0$ and $t > 0$. Setting $q = 0$ yields $\mathbf{P}\{q_{\text{all}}(t) = 0\} \leqslant \mathbf{P}\{q_{L,\text{all}}(t) = 0\}$. When the retrial queue can reach its steady state, the limit $\lim_{t\to\infty} \mathbf{P}\{q_{\text{all}}(t) = 0\}$ exists and is positive. Since the Markov process $\{q_{L,\text{all}}(t), I(t), I_i(t), 1 \leqslant i \leqslant \min\{s, q_{L,\text{all}}(t)\}\}$ of the BMAP/PH/s queue is irreducible, the limit $\lim_{t\to\infty} \mathbf{P}\{q_{L,\text{all}}(t) = 0\}$ exists and is positive since $\lim_{t\to\infty} \mathbf{P}\{q_{L,\text{all}}(t) = 0\} \geqslant \lim_{t\to\infty} \mathbf{P}\{q_{\text{all}}(t) = 0\} > 0$. This implies that the BMAP/PH/s queue can reach its steady state. When the BMAP/PH/s queue is ergodic, $\rho < 1$ must be true [2]. This completes the proof. □

Note. Theorem 2 can be extended to more general retrial queueing systems such as GI/G/s/$s+K$ retrial queues with general retrial times and BMAP/G/s/$s+K$ retrial

queues with general retrial times, as long as the ergodicity of these queueing systems is well defined.

5. Proof of the sufficiency

To prove the sufficiency, a Foster's criterion (or mean-drift method) is utilized [4,12]. The key is to introduce a test function (see equation (5.3)) and to determine a number of constants in the test function so that one of Foster's criteria for ergodicity is satisfied for the test function.

For later use, we first introduce the following subsets of the states of the PH-distribution of the retrial times, which play an important role in the proof of sufficiency. Let $\Im(1) = \{i\colon\ 1 \leqslant i \leqslant m_2,\ h_i^0 \neq 0\}$ and, for $2 \leqslant j \leqslant j_0$,

$$\Im(j) = \left\{ i\colon\ 1 \leqslant i \leqslant m_2,\ i \notin \bigcup_{k=1}^{j-1} \Im(k),\ \text{and for somet } t \in \Im(j-1),\ h_{i,t} > 0 \right\}, \tag{5.1}$$

$$\Re(j) = \bigcup_{k=j}^{j_0} \Im(k), \qquad \Re(1) - \Re(j) = \bigcup_{k=1}^{j-1} \Im(k) \quad \text{and} \quad \Re(1) = \{1, 2, \ldots, m_2\},$$

where j_0 is a positive integer such that $\Im(j_0)$ is not empty and $\bigcup_{k=1}^{j_0} \Im(k) = \{1, 2, \ldots, m_2\}$. Notice that $\Im(1)$ is always nonempty. Intuitively, a retrial for service can occur after one transition if the retrial process of a customer is in one of the states in $\Im(1)$; and a retrial for service can occur after (at least) j transitions if the retrial process of a customer is in one of the states in $\Im(j)$ for $1 \leqslant j \leqslant j_0$. In general, the retrial process of a customer can go from a state in $\Im(j)$ to a state in $\Im(j-1), \Im(j), \ldots, \Im(j_0 - 1)$, or $\Im(j_0)$, but not to any state in $\Im(1), \ldots,$ or $\Im(j-2)$ after one transition. For instance, when the retrial times have an (common) exponential distribution, $\Im(1) = \Re(1) = \{1\}$ and $j_0 = 1$. When the retrial times have

$$H = \begin{pmatrix} -5 & 4 \\ 3 & -3 \end{pmatrix}, \qquad \Im(1) = \{1\}, \qquad \Im(2) = \{2\},$$
$$\Re(1) = \{1, 2\}, \qquad \Re(2) = \{2\} \quad \text{and} \quad j_0 = 2.$$

When $j_0 = 1$, i.e., a customer can retry for service in any state, the proof of sufficiency is a straightforward generalization of Diamond and Alfa [7, proposition 1]. When $j_0 > 1$, the proof becomes much more complicated because of the states in the subsets $\Im(2), \Im(3), \ldots,$ and $\Im(j_0)$. Therefore, the proof of the following theorem is rather long and an outline of the proof is given right after equation (5.3).

Theorem 3. When $\rho < 1$, the Markov process $\{N_i(t),\ 1 \leqslant i \leqslant m_2,\ q(t),\ I(t),\ I_j(t),\ 1 \leqslant j \leqslant \min\{s, q(t)\}\}$ (or, equivalently, the BMAP/PH/s/$s + K$ retrial queue with PH-retrial times defined in section 2) is ergodic.

Proof. According to Foster's criterion (see [12, statement 1]), the Markov process $\{N_i(t),\ 1 \leqslant i \leqslant m_2,\ q(t),\ I(t),\ I_j(t),\ 1 \leqslant j \leqslant \min\{s, q(t)\}\}$ is ergodic if there exists a vector-valued test (or Lyapunov) function $\{\mathbf{f_n},\ \mathbf{n} \in \mathcal{N}\}$ such that $\mathbf{f_n} \to \infty$ when $\sum_{i=1}^{m_2} n_i \to \infty$ and

$$\sum_{i=1}^{m_2} A(\mathbf{n}, \mathbf{n}-\mathbf{e}_i)\mathbf{f}_{\mathbf{n}-\mathbf{e}_i} + A(\mathbf{n}, \mathbf{n})\mathbf{f}_{\mathbf{n}} + \sum_{i=1}^{m_2} \sum_{j=1:\ j\neq i}^{m_2} A(\mathbf{n}, \mathbf{n}-\mathbf{e}_i+\mathbf{e}_j)\mathbf{f}_{\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j} + \sum_{\mathbf{k}\geqslant 0, \mathbf{k}\neq 0} A(\mathbf{n}, \mathbf{n}+\mathbf{k})\mathbf{f}_{\mathbf{n}+\mathbf{k}} \leqslant -\varepsilon\,\mathbf{e} \tag{5.2}$$

holds for all but a finite number of $\mathbf{n} \in \mathcal{N}$ for some positive ε. In the above and the following inequalities, $A(\mathbf{n}, \mathbf{n}') = 0$ if $\mathbf{n}' \notin \mathcal{N}$. Therefore, the key of the proof is to construct such a test function $\{\mathbf{f_n},\ \mathbf{n} \in \mathcal{N}\}$. For $\mathbf{n} \in \mathcal{N}$, let

$$\mathbf{f_n} = a \sum_{j=1}^{j_0} z^{\sum_{i\in\Re(j)} n_i}\mathbf{e} + z^{\sum_{i=1}^{m_2} n_i}\mathbf{u} \equiv \sum_{j=1}^{j_0} \mathbf{f_n}(1, j) + \mathbf{f_n}(2), \tag{5.3}$$

where $1 < z < z_0$, $\mathbf{u}$ is a vector of size M, a is a positive number, and subsets $\{\Re(j)\}$ and j_0 are introduced in equation (5.1). Vectors $\{\mathbf{f_n}(1, j)\}$, $1 \leqslant j \leqslant j_0$, $\{\mathbf{f_n}(2)\}$ and sets $\{\Im(j),\ 1 < j \leqslant j_0\}$ are used as follows. Vector $\mathbf{f_n}(1, 1) + \mathbf{f_n}(2)$ is used to guarantee that inequality (5.2) holds when n_i is large for $i \in \Im(1)$, and vectors $\{\mathbf{f_n}(1, j),\ 2 \leqslant j \leqslant j_0\}$ are used to guarantee that inequality (5.2) holds when n_i is large for $i \in \Im(j)$ and $2 \leqslant j \leqslant j_0$. The difficult part is to make the last mm_1^s elements of the left hand side of inequality (5.2) negative, which is achieved by using results associated with the classical BMAP/PH/s queue with $\rho < 1$. Values of parameters a, $\mathbf{u}$ and z must be determined so that inequality (5.2) holds for all but a finite number of $\mathbf{n} \in \mathcal{N}$ for some positive ε. This is done in several steps.

Step 1. In order to determine the constants $\{a, \mathbf{u}, z\}$, we evaluate equation (5.2) when function $\mathbf{f_n}$ is replaced by its expression (5.3).

Step 2. With the explicit expression (5.8) obtained from step 1, we argue that equation (5.2) holds if all the constants can be chosen appropriately. In this step, we first argue that equation (5.2) holds when $\Im(1) = \{1, 2, \ldots, m_2\}$. Then we generalize the conclusion to cases with $j_0 \geqslant 2$.

Step 3. Finally, we determine these constants explicitly. The constants are obtained by using results related to the classical BMAP/PH/s queue.

Step 1. The evaluation of the left hand side of inequality (5.2) is a complicated process and is divided into several steps. First, terms containing vectors $\{\mathbf{f_n}(1, 1)\}$ on the left-hand side of inequality (5.2) are evaluated.

$$
\begin{aligned}
&\sum_{i=1}^{m_2} A(\mathbf{n}, \mathbf{n}-\mathbf{e}_i)\mathbf{f}_{\mathbf{n}-\mathbf{e}_i}(1,1) + A(\mathbf{n},\mathbf{n})\mathbf{f}_{\mathbf{n}}(1,1) \\
&\quad + \sum_{i=1}^{m_2}\sum_{j=1:\ j\neq i}^{m_2} A(\mathbf{n}, \mathbf{n}-\mathbf{e}_i+\mathbf{e}_j)\mathbf{f}_{\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j}(1,1) + \sum_{\mathbf{k}\geqslant 0,\,\mathbf{k}\neq 0} A(\mathbf{n},\mathbf{n}+\mathbf{k})\mathbf{f}_{\mathbf{n}+\mathbf{k}}(1,1) \\
&= z^{\sum_{i=1}^{m_2} n_i - 1}\left[\sum_{i=1}^{m_2} A(\mathbf{n},\mathbf{n}-\mathbf{e}_i) + zA(\mathbf{n},\mathbf{n}) + z\sum_{i=1}^{m_2}\sum_{j=1:\ j\neq i}^{m_2} A(\mathbf{n},\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j)\right. \\
&\qquad \left. + \sum_{\mathbf{k}\geqslant 0,\,\mathbf{k}\neq 0} z^{1+\sum_{i=1}^{m_2} k_i} A(\mathbf{n},\mathbf{n}+\mathbf{k})\right] a\mathbf{e} \\
&= z^{\sum_{i=1}^{m_2} n_i - 1}\left[\left(\sum_{i=1}^{m_2} n_i h_i^0\right)A_2 + zA_1 + z\sum_{i=1}^{m_2} n_i h_{i,i}\mathbf{I} + z\sum_{i=1}^{m_2} n_i h_i^0(\mathbf{I}-\Gamma)\right. \\
&\qquad \left. + z\sum_{i=1}^{m_2}\sum_{j=1:\ j\neq i}^{m_2} n_i h_{i,j}\mathbf{I} + \sum_{\mathbf{k}\geqslant 0,\,\mathbf{k}\neq 0} z^{1+\sum_{i=1}^{m_2} k_i} p(\mathbf{k},\beta)A_0\left(\sum_{j=1}^{m_2} k_j\right)\right] a\mathbf{e} \\
&= z^{\sum_{i=1}^{m_2} n_i - 1}\left\{\left(\sum_{i=1}^{m_2} n_i h_i^0\right)A_2 + zA_1 + z\sum_{i=1}^{m_2} n_i h_{i,i}\mathbf{I} + z\sum_{i=1}^{m_2} n_i h_i^0(\mathbf{I}-\Gamma)\right. \\
&\qquad \left. + z\sum_{i=1}^{m_2}\sum_{j=1:\ j\neq i}^{m_2} n_i h_{i,j}\mathbf{I} + \sum_{N=1}^{\infty} z^{1+N}\left[\left(\sum_{\mathbf{k}:\ \sum_{j=1}^{m_2} k_j = N} p(\mathbf{k},\beta)\right)A_0(N)\right]\right\} a\mathbf{e} \\
&= z^{\sum_{i=1}^{m_2} n_i - 1}\left\{zA_1 + z\sum_{N=1}^{\infty} z^N A_0(N) + \left(\sum_{i=1}^{m_2} n_i h_i^0\right)A_2 + z\sum_{i=1}^{m_2} n_i h_{i,i}\mathbf{I}\right. \\
&\qquad \left. + z\sum_{i=1}^{m_2} n_i h_i^0(\mathbf{I}-\Gamma) + z\sum_{i=1}^{m_2}\sum_{j=1:\ j\neq i}^{m_2} n_i h_{i,j}\mathbf{I}\right\} a\mathbf{e} \\
&= z^{\sum_{i=1}^{m_2} n_i - 1}\left\{a\sum_{N=1}^{\infty} z\left(z^N - 1\right)A_0(N)\mathbf{e}\right. \\
&\qquad \left. + a\left[\sum_{i=1}^{m_2} n_i\left(h_i^0 A_2 + zh_{i,i}\mathbf{I} + zh_i^0(\mathbf{I}-\Gamma) + z\sum_{j=1:\ j\neq i}^{m_2} h_{i,j}\mathbf{I}\right)\right]\mathbf{e}\right\} \\
&= z^{\sum_{i=1}^{m_2} n_i - 1}\left\{a\sum_{N=1}^{\infty} z\left(z^N - 1\right)A_0(N)\mathbf{e} + a\left(\sum_{i=1}^{m_2} n_i h_i^0\right)(A_2 - z\Gamma)\mathbf{e}\right\}. \qquad (5.4)
\end{aligned}
$$

The last three equalities hold because the total sum of the probabilities of a multinomial distribution equals one, $[\sum_{N\geqslant 1} A_0(N) + A_1]\mathbf{e} = 0$, and $T\mathbf{e} + \mathbf{T}^0 = 0$, i.e., $\sum_j h_{i,j} + h_i^0 = 0$ for every i, respectively. Similarly, terms containing vectors $\{\mathbf{f}_{\mathbf{n}}(2)\}$ on left-hand side of inequality (5.2) can be evaluated. The result is given as

$$\sum_{i=1}^{m_2} A(\mathbf{n}, \mathbf{n}-\mathbf{e}_i)\mathbf{f}_{\mathbf{n}-\mathbf{e}_i}(2) + A(\mathbf{n}, \mathbf{n})\mathbf{f}_{\mathbf{n}}(2) + \sum_{i=1}^{m_2}\sum_{j=1:\ j\neq i}^{m_2} A(\mathbf{n}, \mathbf{n}-\mathbf{e}_i+\mathbf{e}_j)\mathbf{f}_{\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j}(2)$$
$$+ \sum_{\mathbf{k}\geqslant 0,\,\mathbf{k}\neq 0} A(\mathbf{n}, \mathbf{n}+\mathbf{k})\mathbf{f}_{\mathbf{n}+\mathbf{k}}(2)$$
$$= z^{\sum_{i=1}^{m_2} n_i - 1}\left\{ z\left[A_1 + \sum_{N=1}^{\infty} z^N A_0(N)\right]\mathbf{u} + \left(\sum_{i=1}^{m_2} n_i h_i^0\right)(A_2 - z\Gamma)\mathbf{u}\right\}. \tag{5.5}$$

For $2 \leqslant j \leqslant j_0$, terms containing vectors $\{\mathbf{f}_{\mathbf{n}}(1, j)\}$ on the left-hand side of inequality (5.2) are evaluated as follows:

$$\sum_{i=1}^{m_2} A(\mathbf{n}, \mathbf{n}-\mathbf{e}_i)\mathbf{f}_{\mathbf{n}-\mathbf{e}_i}(1, j) + A(\mathbf{n}, \mathbf{n})\mathbf{f}_{\mathbf{n}}(1, j)$$
$$+ \sum_{i=1}^{m_2}\sum_{t=1:\ t\neq i}^{m_2} A(\mathbf{n}, \mathbf{n}-\mathbf{e}_i+\mathbf{e}_t)\mathbf{f}_{\mathbf{n}-\mathbf{e}_i+\mathbf{e}_t}(1, j) + \sum_{\mathbf{k}\geqslant 0,\,\mathbf{k}\neq 0} A(\mathbf{n}, \mathbf{n}+\mathbf{k})\mathbf{f}_{\mathbf{n}+\mathbf{k}}(1, j)$$
$$= a\left\{ \sum_{i=1}^{m_2} n_i h_i^0 A_2 z^{\sum_{t\in\Re(j)} n_t} + \left[A_1 + \sum_{i=1}^{m_2} n_i h_{i,i}\mathbf{I} + \sum_{i=1}^{m_2} n_i h_i^0(\mathbf{I}-\Gamma)\right] z^{\sum_{t\in\Re(j)} n_t}\right.$$
$$+ \left(\sum_{i=1:\ i\notin\Re(j)}^{m_2}\ \sum_{t=1:\ t\neq i,\ t\notin\Re(j)}^{m_2} n_i h_{i,t}\mathbf{I}\right) z^{\sum_{t\in\Re(j)} n_t}$$
$$+ \left(\sum_{i=1:\ i\notin\Re(j)}^{m_2}\ \sum_{t=1:\ t\neq i,\ t\in\Re(j)}^{m_2} n_i h_{i,t}\mathbf{I}\right) z^{1+\sum_{t\in\Re(j)} n_t}$$
$$+ \left(\sum_{i=1:\ i\in\Re(j)}^{m_2}\ \sum_{t=1:\ t\notin\Re(j)}^{m_2} n_i h_{i,t}\mathbf{I}\right) z^{\sum_{t\in\Re(j)} n_t - 1}$$
$$+ \left(\sum_{i=1:\ i\in\Re(j)}^{m_2}\ \sum_{t=1:\ t\neq i,\ t\in\Re(j)}^{m_2} n_i h_{i,t}\mathbf{I}\right) z^{\sum_{t\in\Re(j)} n_t}$$
$$+ \sum_{\mathbf{k}:\ k_t=0,\ t\in\Re(j)}\left[p(\mathbf{k}, \beta)A_0\left(\sum_{i=1}^{m_2} k_i\right) z^{\sum_{t\in\Re(j)} n_t}\right]$$
$$\left.+ \sum_{\mathbf{k}:\ \exists k_t>0,\ t\in\Re(j)}\left[p(\mathbf{k}, \beta)A_0\left(\sum_{i=1}^{m_2} k_i\right) z^{\sum_{t\in\Re(j)}(n_t+k_t)}\right]\right\}\mathbf{e}$$
$$= a z^{\sum_{t\in\Re(j)} n_t - 1}\left\{ z \sum_{\mathbf{k}:\ \exists k_t>0,\ t\in\Re(j)}\left[p(\mathbf{k}, \beta)A_0\left(\sum_{i=1}^{m_2} k_i\right)\left(z^{\sum_{t\in\Re(j)} k_t} - 1\right)\right]\mathbf{e}\right.$$
$$\left.+ z(z-1)\sum_{i\notin\Re(j)} n_i\left(\sum_{t\in\Re(j)} h_{i,t}\right)\mathbf{e} - (z-1)\sum_{i\in\Re(j)} n_i\left(\sum_{t\notin\Re(j)} h_{i,t}\right)\mathbf{e}\right\}, \tag{5.6}$$

in which $[\sum_{N\geqslant 1} A_0(N) + A_1]\mathbf{e} = 0$, $T\mathbf{e} + \mathbf{T}^0 = 0$, i.e., $\sum_j h_{i,j} + h_i^0 = 0$, and $A_2\mathbf{e} = \Gamma\mathbf{e}$ are used. The notation $\exists$ stands for "there exists". Also notice that $h_i^0 = 0$ for $i \in \Re(j)$ if $j \geqslant 2$. Summing up equations (5.4)–(5.6) from $j = 2$ to $j = j_0$, the left-hand side of inequality (5.2) becomes

$$
\begin{aligned}
& z^{\sum_{i=1}^{m_2} n_i - 1}\Bigg\{ a\sum_{N=1}^{\infty} z\big(z^N - 1\big)A_0(N)\mathbf{e} - a(z-1)\Bigg(\sum_{i=1}^{m_2} n_i h_i^0\Bigg)\Gamma\mathbf{e} \\
& \quad + z\Bigg[A_1 + \sum_{N=1} z^N A_0(N)\Bigg]\mathbf{u} + \Bigg(\sum_{i=1}^{m_2} n_i h_i^0\Bigg)(A_2 - z\Gamma)\mathbf{u}\Bigg\} \\
& \quad + a\sum_{j=2}^{j_0} z^{\sum_{i\in\Re(j)} n_i - 1}\Bigg\{ z \sum_{\mathbf{k}:\, \exists k_t > 0,\ t\in\Re(j)} \Bigg[p(\mathbf{k},\beta)A_0\Bigg(\sum_{i=1}^{m_2} k_i\Bigg)\Big(z^{\sum_{i\in\Re(j)} k_i} - 1\Big)\Bigg]\mathbf{e} \\
& \quad + z(z-1)\sum_{i\notin\Re(j)} n_i\Bigg(\sum_{t\in\Re(j)} h_{i,t}\Bigg)\mathbf{e} - (z-1)\sum_{i\in\Re(j)} n_i\Bigg(\sum_{t\notin\Re(j)} h_{i,t}\Bigg)\mathbf{e}\Bigg\}. \qquad (5.7)
\end{aligned}
$$

Rearranging terms in the third and fourth lines in equation (5.7) with respect to the subsets $\{\Im(j),\ 1 \leqslant j \leqslant j_0\}$, inequality (5.2) becomes

$$
\begin{aligned}
& z^{\sum_{i=1}^{m_2} n_i - 1}\Bigg\{ a\sum_{N=1}^{\infty} z\big(z^N - 1\big)A_0(N)\mathbf{e} - a(z-1)\Bigg(\sum_{i=1}^{m_2} n_i h_i^0\Bigg)\Gamma\mathbf{e} \\
& \qquad + z\Bigg[A_1 + \sum_{N=1} z^N A_0(N)\Bigg]\mathbf{u} + \Bigg(\sum_{i=1}^{m_2} n_i h_i^0\Bigg)(A_2 - z\Gamma)\mathbf{u}\Bigg\} \\
& \quad + a(z-1)\sum_{i\in\Im(1)} n_i\Bigg[z\sum_{t=2}^{j_0} z^{\sum_{l\in\Re(t)} n_l - 1}\Bigg(\sum_{l\in\Re(t)} h_{i,l}\Bigg)\Bigg]\mathbf{e} \\
& \quad + a\sum_{j=2}^{j_0}\Bigg\{ z^{\sum_{i\in\Re(j)} n_i} \sum_{\mathbf{k}:\, \exists k_t > 0,\ t\in\Re(j)} \Bigg[p(\mathbf{k},\beta)A_0\Bigg(\sum_{i=1}^{m_2} k_i\Bigg)\Big(z^{\sum_{i\in\Re(j)} k_i} - 1\Big)\Bigg]\mathbf{e} \\
& \qquad + (z-1)\sum_{i\in\Im(j)} n_i\Bigg[\sum_{t=j+1}^{j_0} z^{\sum_{l\in\Re(t)} n_l}\Bigg(\sum_{l\in\Re(t)} h_{i,l}\Bigg) \\
& \qquad - z^{\sum_{l\in\Re(j)} n_l - 1}\sum_{t\in\Im(j-1)} h_{i,t}\Bigg]\mathbf{e}\Bigg\} \leqslant -\varepsilon\mathbf{e} \qquad (5.8)
\end{aligned}
$$

by noticing that

$$
\sum_{i\in\Re(j)} n_i\Bigg(\sum_{t\notin\Re(j)} h_{i,t}\Bigg) = \sum_{i\in\Im(j)} n_i\Bigg(\sum_{t\in\Im(j-1)} h_{i,t}\Bigg)
$$

and the definition of $\Re(j)$ and $\Im(j)$ given in equation (5.1).

Step 2. We now argue that when parameters a, $\mathbf{u}$ and z can be chosen properly, inequality (5.8), or equivalently inequality (5.2), holds for all but a finite number of $\mathbf{n} \in \mathcal{N}$ for some positive ε.

We begin with the first line of inequality (5.8), which is related to vectors $\{\mathbf{f_n}(1,1)\}$ of the test function. Since $D^*(z) = \sum_{n=0}^{\infty} z^n D_n$ is finite for $1 < z < z_0$, $\sum_{n=1}^{\infty} z^n D_{n+t}$ is finite and uniformly bounded by $D^*(z) - D_0$ for all $t \geqslant 0$. Then, $\sum_{n=1}^{\infty} z(z^n - 1)A_0(N)$ is finite for any fixed z, $1 < z < z_0$. The value of z shall be specified later. Thus, for any fixed $1 < z < z_0$ and before multiplying the term $z^{n_1+\cdots+n_{m_2}-1}$, the first line of inequality (5.8), except its last mm_1^s elements, becomes negative if at least one value in $\{n_i\colon\ i \in \Im(1)\}$ is large enough. It is clear that the last mm_1^s elements of the first line of inequality (5.8) are nonnegative, finite, and independent of state $\mathbf{n}$ before multiplying the term $z^{n_1+\cdots+n_{m_2}-1}$. In order to make inequality (5.8) true for all but a finite number of $\mathbf{n} \in \mathcal{N}$ for some positive ε, we need to make the last mm_1^s elements negative. This leads to the second line of inequality (5.8), which is related to vectors $\{\mathbf{f_n}(2)\}$ of the test function.

For the second line of inequality (5.8), we choose a positive vector $\mathbf{u}$ such that $(A_2 - z\Gamma)\mathbf{u} = 0$ and the last mm_1^s elements of $(A_1 + \sum_{N=1}^{\infty} z^N A_0(N))\mathbf{u}$ are negative. Such a positive vector $\mathbf{u}$ exists when $\rho < 1$ and z is close to 1 ($z > 1$), which shall be specified later. Also notice that the selection of $\mathbf{u}$ is independent of level $\mathbf{n}$.

Next, we consider the first line and the second line of inequality (5.8) together. Since the last mm_1^s elements of

$$a \sum_{N=1}^{\infty} z\big(z^N - 1\big) A_0(N)\mathbf{e}$$

are

$$a \sum_{N=1}^{\infty} z\big(z^N - 1\big)(D_N \otimes \mathbf{I}_{m_1^s})\mathbf{e}$$

and $D^*(z)$ is finite, a small a can be chosen so that the last mm_1^s elements of

$$\left(A_1 + \sum_{N=1}^{\infty} z^N A_0(N)\right)\mathbf{u} + a \sum_{N=1}^{\infty} z\big(z^N - 1\big) A_0(N)\mathbf{e}$$

are negative. Then the last mm_1^s elements of the sum of the first and second lines of inequality (5.8) are negative for any fixed z and its corresponding vector $\mathbf{u}$, where z is close to 1 ($z > 1$). This implies that all elements of the sum of the first and second lines of inequality (5.8), before multiplying the term $z^{n_1+\cdots+n_{m_2}-1}$, are less than $-\varepsilon$ for some positive ε if at least one value in $\{n_i\colon\ i \in \Im(1)\}$ is large enough. Therefore, all elements of the sum of the first and second lines of inequality (5.8) are less than $-\varepsilon z^{n_1+\cdots+n_{m_2}-1}$ for some positive ε if at least one value in $\{n_i\colon\ i \in \Im(1)\}$ is large enough.

It follows from the above argument that if $\Im(1) = \{1, 2, \ldots, m_2\}$, i.e., a retrial may occur in any state of the retrial process of an orbiting customer, inequality (5.2) holds for all but a finite number of $\mathbf{n} \in \mathcal{N}$ for some positive ε, provided that an appropriate vector $\mathbf{u}$ can be found. However, it is possible that $\Im(1) \neq \{1, 2, \ldots, m_2\}$ for a PH-distribution. Thus, vectors $\{\mathbf{f_n}(1, j)\}$, $2 \leqslant j \leqslant j_0$, are included in the test function to deal with the case when none in $\{n_i\colon\ i \in \Im(1)\}$ is large and some value in $\{n_i\colon\ i \notin \Im(1)\}$ is large. This leads to the third, fourth, and fifth lines on the left hand side of inequality (5.8).

Now, we consider the third line of inequality (5.8). Based on the above discussion, if at least one value in $\{n_i\colon\ i \in \Im(1)\}$ is large, the sum of the first, second, and third lines of inequality (5.8) is less than

$$
\begin{aligned}
& z^{\sum_{i=1}^{m_2} n_i - 1} \left\{ -\varepsilon \mathbf{e} + a(z-1) \sum_{i \in \Im(1)} n_i \left[z \sum_{t=2}^{j_0} z^{-\sum_{l \in \Re(1),\ l \notin \Re(t)} n_l} \sum_{l \in \Re(t)} h_{i,l} \right] \mathbf{e} \right\} \\
& \quad \leqslant z^{\sum_{i=1}^{m_2} n_i - 1} \left\{ -\varepsilon \mathbf{e} + az(z-1) \sum_{i \in \Im(1)} \frac{n_i}{z^{\sum_{l \in \Re(1)} n_l - 1}} \left[\sum_{t=2}^{j_0} \sum_{l \in \Re(t)} h_{i,l} \right] \mathbf{e} \right\} \\
& \quad \leqslant z^{\sum_{i=1}^{m_2} n_i - 1} \left\{ -\varepsilon \mathbf{e} + \sum_{i \in \Im(1)} \frac{az^2(z-1) n_i}{1 + 0.5 \left(\sum_{l \in \Im(1)} n_l \right)^2 (z-1)^2} \left[\sum_{t=2}^{j_0} \sum_{l \in \Re(t)} h_{i,l} \right] \mathbf{e} \right\} \\
& \quad \leqslant -z^{\sum_{i=1}^{m_2} n_i - 1} (\varepsilon - \delta) \mathbf{e}, \qquad (5.9)
\end{aligned}
$$

where δ can be arbitrarily small. Notice that $\Im(1) \subseteq \{i\colon\ i \notin \Re(j)\}$ when

$$
j > 1 \quad \text{and} \quad z^{n+1} = \left[1 + (z-1) \right]^{n+1} \geqslant 1 + \frac{n(n+1)(z-1)^2}{2} \geqslant 1 + 0.5 n^2 (z-1)^2.
$$

Thus, the sum of the first, second, and third lines of inequality (5.8) is less than $-\varepsilon z^{n_1 + \cdots + n_{m_2} - 1}$ for some positive ε if at least one value in $\{n_i\colon\ i \in \Im(1)\}$ is large enough. (Note that, to simplify the notation, we replace $\varepsilon - \delta$ by ε in the last line of inequality (5.9). Similar substitution shall take place in inequality (5.11) and in the discussion following equation (5.12).)

The last three lines of the left-hand side of inequality (5.8) can be rewritten as $(z > 1)$

$$
\begin{aligned}
& a \sum_{j=2}^{j_0} z^{\sum_{i \in \Re(j)} n_i - 1} \left\{ z \sum_{\mathbf{k}\colon\ \exists k_t > 0,\ t \in \Re(j)} \left[p(\mathbf{k}, \beta) A_0 \left(\sum_{i=1}^{m_2} k_i \right) \left(z^{\sum_{i \in \Re(j)} k_i} - 1 \right) \right] \right. \\
& \qquad \left. + (z-1) \sum_{i \in \Im(j)} n_i \left[z \sum_{t=j+1}^{j_0} z^{-\sum_{l \notin \Re(t),\ l \in \Re(j)} n_l} \sum_{l \in \Re(t)} h_{i,l} - \sum_{t \in \Im(j-1)} h_{i,t} \right] \right\} \mathbf{e}
\end{aligned}
$$

$$\leqslant \sum_{j=2}^{j_0} z^{\sum_{i\in\Re(j)} n_i - 1} a \Bigg\{ z \sum_{N=1}^{\infty} \big(z^N - 1\big) A_0(N) + \sum_{i\in\Im(j)} \frac{(z-1) z n_i}{z^{\sum_{l\in\Re(j)} n_l}} \Bigg(\sum_{t=j+1}^{j_0} \sum_{l\in\Re(t)} h_{i,l} \Bigg) - (z-1) \sum_{i\in\Im(j)} n_i \sum_{t\in\Im(j-1)} h_{i,t} \Bigg\} \mathbf{e}$$

$$\equiv \sum_{j=2}^{j_0} z^{\sum_{i\in\Re(j)} n_i} \Delta(j, \mathbf{n}). \tag{5.10}$$

By definition, there must be at least one positive $h_{i,t}$ in $\{h_{i,t},\ t \in \Im(j-1)\}$ for every $i \in \Im(j)$, i.e., $\sum_{t\in\Im(j-1)} h_{i,t} > 0$. Since $z > 1$, $n_i z^{-\sum_{l\in\Re(j)} n_l}$ is a bounded function for every $i \in \Im(j)$. Then, for $2 \leqslant j \leqslant j_0$, $\Delta(j, \mathbf{n})$ becomes negative when at least one value in $\{n_i\colon\ i \in \Im(j)\}$ is large enough. It follows that vector $\Delta(j, \mathbf{n})$ is uniformly bounded from above with respect to $\mathbf{n} \in \mathcal{N}$ and j (for any fixed z).

Combining inequalities (5.9) and (5.10) together, we obtain that, if at least one value in $\{n_i\colon\ i \in \Im(1)\}$ is large, the left-hand side of inequality (5.8) is less than

$$z^{\sum_{i=1}^{m_2} n_i - 1} \Bigg\{ -\varepsilon \mathbf{e} + \sum_{j=2}^{j_0} \frac{\Delta(j, \mathbf{n})}{z^{\sum_{i\in\Re(j)} n_i}} \Bigg\} \leqslant z^{\sum_{i=1}^{m_2} n_i - 1} \Bigg\{ -\varepsilon \mathbf{e} + \frac{1}{z^{\sum_{i\in\Im(1)} n_i}} \sum_{j=2}^{j_0} \Delta(j, \mathbf{n}) \Bigg\}. \tag{5.11}$$

Since functions $\{\Delta(j, \mathbf{n}),\ 2 \leqslant j \leqslant j_0\}$ are uniformly bounded from above with respect to $\mathbf{n} \in \mathcal{N}$, it is easy to see that the last expression in inequality (5.11) is less than $-\varepsilon z^{n_1 + \cdots + n_{m_2} - 1} \mathbf{e}$ for some positive ε when at least one value in $\{n_i\colon\ i \in \Im(1)\}$ is large enough. This implies that the left-hand side of inequality (5.8) is less than $-\varepsilon z^{n_1 + \cdots + n_{m_2} - 1} \mathbf{e}$ for some positive ε if at least value in $\{n_i\colon\ i \in \Im(1)\}$ is large enough, regardless of the values in $\{n_i\colon\ i \notin \Im(1)\}$. Let $\mathcal{N}(1)$ be the subset of $\mathcal{N}$ such that the left-hand side of inequality (5.8) is less than $-\varepsilon z^{n_1 + \cdots + n_{m_2} - 1} \mathbf{e}$ for some positive ε if at least one value in $\{n_i\colon\ i \in \Im(1)\}$ is large enough.

To complete the proof, we still need to show that (5.2) holds when no value in $\{n_i\colon\ i \in \Im(1)\}$ is large and some value in $\{n_i\colon\ i \notin \Im(1)\}$ is large. Similar to inequality (5.10), we denote the sum of the first three lines of inequality (5.8) as $z^{n_1 + \cdots + n_{m_2} - 1} \Phi(\mathbf{n})$. It is easy to see that $\Phi(\mathbf{n})$ is uniformly bounded from above with respect to $\mathbf{n}$ (see inequality (5.9)). Rewrite inequality (5.8) as

$$z^{\sum_{i\in\Re(2)} n_i - 1} \Bigg[z^{\sum_{i\in\Im(1)} n_i} \Phi(\mathbf{n}) + \Delta(2, \mathbf{n}) + \sum_{j=3}^{j_0} \frac{\Delta(j, \mathbf{n})}{z^{\sum_{i\in\Re(2),\ i\notin\Re(j)} n_i}} \Bigg]. \tag{5.12}$$

Since $z^{\sum_{i\in\Im(1)} n_i}$ is uniformly bounded from above for $\mathbf{n}$ in $\mathcal{N} - \mathcal{N}(1)$ and $\Phi(\mathbf{n})$ is uniformly bounded from above for $\mathbf{n}$ in $\mathcal{N}$, function $z^{\sum_{i\in\Im(1)} n_i} \Phi(\mathbf{n})$ is uniformly bounded from above for $\mathbf{n}$ in $\mathcal{N} - \mathcal{N}(1)$. Thus, expression (5.12) is less than $-z^{\sum_{i\in\Re(2)} n_i - 1} \varepsilon \mathbf{e}$ for some positive ε for $\mathbf{n}$ in $\mathcal{N} - \mathcal{N}(1)$ and at least one value in $\{n_i\colon\ i \in \Im(2)\}$ is

large enough, regardless of the values in $\{n_i\colon\ i \notin \Im(2)\}$. Let $\mathcal{N}(2)$ be the subset of $\mathcal{N} - \mathcal{N}(1)$ such that the left hand side of inequality (5.8) is less than $-z^{\sum_{i\in\Re(2)} n_i - 1}\varepsilon\mathbf{e}$ for some positive ε if at least one value in $\{n_i\colon\ i \in \Im(2)\}$ is large enough, regardless of the values in $\{n_i\colon\ i \notin \Im(2)\}$. Similarly, for $3 \leqslant j \leqslant j_0$, we can find $\mathcal{N}(j)$, a subset of $\mathcal{N} - \mathcal{N}(1) \cup \cdots \cup \mathcal{N}(j-1)$ such that for $\mathbf{n}$ in $\mathcal{N}(j)$ the left-hand side of inequality (5.8) is less than $-z^{\sum_{i\in\Re(j)} n_i - 1}\varepsilon\mathbf{e}$ for some positive ε if at least one value in $\{n_i\colon\ i \in \Im(j)\}$ is large enough, regardless of the values of $\{n_i\colon\ i \notin \Im(j)\}$. It is easy to see that for any $\mathbf{n}$ in $\mathcal{N}(1) \cup \cdots \cup \mathcal{N}(j_0)$, inequality (5.3) holds for some positive ε. Since $\mathcal{N} - \mathcal{N}(1) \cup \cdots \cup \mathcal{N}(j_0)$ has only a finite number of members, we have proved that inequality (5.2) holds for all but a finite number of $\mathbf{n} \in \mathcal{N}$ for some positive ε, provided that a positive vector $\mathbf{u}$ can be found. To see why $\mathcal{N} - \mathcal{N}(1) \cup \cdots \cup \mathcal{N}(j_0)$ has a finite number of members, consider the special case with $m_2 = 2$, $\Im(1) = \{1\}$ and $\Im(2) = \{2\}$.

Step 3. Finally, we determine vector $\mathbf{u}$ and the values of other parameters. One of the choices of $\mathbf{u}$ has the following structure:

$$\mathbf{u} = \begin{pmatrix} W_{s+K} \\ \vdots \\ W_1 \end{pmatrix} \mathbf{v}, \tag{5.13}$$

where W_i is an $mm_1^k \times mm_1^s$ nonnegative matrix with $k = \min\{s, i\}$ for $1 \leqslant i \leqslant s+K$, and $\mathbf{v}$ is a positive vector of size mm_1^s. To determine vector $\mathbf{v}$, we consider the classical BMAP/PH/s queue.

When $\rho < 1$, the BMAP/PH/s queue is ergodic [2], which means that the corresponding quasi-birth-and-death (QBD) Markov process $\{q(t),\ I(t),\ I_i(t),\ 1 \leqslant i \leqslant \min\{s, q(t)\}\}$ of the MAP/PH/s queue is ergodic. When $q(t) > s$, the transition blocks of the QBD Markov process are $\{B_{s+1,s}, B_{s+1,s+1}, B_{s+1,s+2}, \ldots\}$ (see section 2 for definitions and extend the definition of $B_{s+1,n}$ to $n > s + K$). Let $B^*(z) = B_{s+1,s} + zB_{s+1,s+1} + z^2B_{s+1,s+2} + \cdots$, for $1 < z < z_0$. Let vector $\mathbf{y}$ be the unique solution to equations $\mathbf{y}B^*(1) = \mathbf{y}(B_{s+1,s} + B_{s+1,s+1} + B_{s+1,s+2} + \cdots) = 0$ and $\mathbf{ye} = 1$. Vector $\mathbf{y}$ is positive since $B^*(1)$ is irreducible. It can be verified that

$$\mathbf{y} = \theta \otimes \left(-\mu\alpha T^{-1}\right) \otimes \left(-\mu\alpha T^{-1}\right) \otimes \cdots \otimes \left(-\mu\alpha T^{-1}\right). \tag{5.14}$$

Then it is easy to verify that $\mathbf{y}B_{s+1,s}\mathbf{e} = s\mu$ and $\mathbf{y}(B_{s+1,s+2} + 2B_{s+1,s+3} + \cdots)\mathbf{e} = \lambda$. Thus, $\rho < 1$ implies that Neuts' condition $\mathbf{y}(B_{s+1,s+2} + 2B_{s+1,s+3} + \cdots)\mathbf{e} = \lambda < s\mu = \mathbf{y}B_{s+1,s}\mathbf{e}$ is satisfied. Denote by $\mathrm{sp}(B^*(z))$ the eigenvalue with the largest real part of $B^*(z)$. Then $\mathrm{sp}(B^*(1)) = 0$. Similar to the proof of Neuts [19, lemma 1.3.3], it can be proved that the derivative of $\mathrm{sp}(B^*(z))$ at $z = 1$ is negative. Thus, $\mathrm{sp}(B^*(z)) < 0$ for z close to 1 and $1 < z < z_0$.

Choose z such that $1 < z < z_0$, z is close to 1, and $\mathrm{sp}(B^*(z)) < 0$. $B^*(z)$ is an irreducible M-matrix (see [13]). Choose $\mathbf{v}$ to be the right eigenvector corresponding to $\mathrm{sp}(B^*(z))$ with the first element to be one. Then $\mathbf{v}$ is positive and satisfies $B^*(z)\mathbf{v} =$

$\mathrm{sp}(B^*(z))\mathbf{v}$, $\mathbf{v} > 0$ and $v_1 = 1$ (note $\mathbf{v}$ is the transpose of vector $(v_1, v_2, \ldots, v_{mm_1^s})$). Based on the special structure of A_2, choose $W_{s+K} = \mathbf{I}$, and

$$W_i = \begin{cases} (\mathbf{I}_{mm_1^i} \otimes \alpha)\dfrac{W_{i+1}}{z}, & 0 \leqslant i \leqslant s-1, \\ \dfrac{W_{i+1}}{z}, & s \leqslant i \leqslant s+K-1. \end{cases} \tag{5.15}$$

It can be verified that every element of vector $\mathbf{u}$ is positive, $(A_2 - z\Gamma)\mathbf{u} = 0$, the last mm_1^s elements of $(A_1 + \sum_{N=1}^{\infty} z^N A_0(N))\mathbf{u}$ are given by $\mathrm{sp}(B^*(z))\mathbf{v}$. When the number of waiting positions K is positive ($K > 0$),

$$B^*(z)\mathbf{v} = \big(B_{s+1,s} + zB_{s+1,s+1} + z^2 B_{s+1,s+2} + \cdots\big)\mathbf{v} = \mathrm{sp}\big(B^*(z)\big)\mathbf{v} < 0.$$

When $K = 0$,

$$\begin{aligned} B^*(z)\mathbf{v} &= \big[B_{s,s-1}(\mathbf{I} \otimes \alpha) + zB_{s,s} + z^2 B_{s,s+1} + \cdots\big]\mathbf{v} \\ &= \big(B_{s+1,s} + zB_{s+1,s+1} + z^2 B_{s+1,s+2} + \cdots\big)\mathbf{v} = \mathrm{sp}\big(B^*(z)\big)\mathbf{v} < 0. \end{aligned}$$

Thus, vector $\mathbf{u}$ obtained from equations (5.13) and (5.15) satisfies our needs. This completes the proof. □

Combining theorems 2 and 3 yields theorem 1. Notice that neither the necessary condition nor the sufficient condition has a direct relationship with the PH-distribution of retrial times. Some intuition on why the BMAP/PH/s/$s + K$ retrial queue with PH-retrial times and the BMAP/PH/s queue have the same ergodicity condition shall be offered in section 8.

6. Ergodicity of retrial queues with impatient customers

There are a number of variations of the retrial queueing system defined in section 2. Among them are the retrial queueing systems with impatient (nonpersistent) customers. In this section, theorem 1 is extended to retrial queueing systems with impatient customers.

Retrial queues with customer loss at arrival epochs. Consider a BMAP/PH/s/$s+K$ retrial queueing system with PH-retrial times and impatient customers. When a customer finds no server and no waiting position available upon arrival (from outside), the customer enters the orbit with probability p and leaves the queueing system with probability $1-p$, $0 \leqslant p \leqslant 1$. Once a customer enters the queueing system, the customer will not leave the system until its service is complete. Thus, the only difference between this retrial queue and the one defined in section 2 occurs at customer arrival epochs.

Theorem 4. The BMAP/PH/s/$s + K$ retrial queueing system with PH-retrial times and customer loss at arrival epochs is ergodic if and only if $\rho = (p\lambda)/(s\mu) < 1$.

Proof. Introduce the Markov process $\{N_i(t),\ 1 \leqslant i \leqslant m_2,\ q(t),\ I(t),\ I_j(t),\ 1 \leqslant j \leqslant \min\{s, q(t)\}\}$ similar to that in section 3. The infinitesimal generator of this Markov process is the same as the one given in section 3 except that

$$A(\mathbf{n},\mathbf{n}+\mathbf{k}) = \begin{pmatrix} 0 & \dots & 0 & \sum_{N=\hat{K}(\mathbf{k})}^{\infty} p(N,\mathbf{k},\beta) D_{s+K+N} \otimes \alpha \otimes \alpha \otimes \cdots \otimes \alpha \\ 0 & \dots & 0 & \sum_{N=\hat{K}(\mathbf{k})}^{\infty} p(N,\mathbf{k},\beta) D_{s-1+K+N} \otimes \mathbf{I}_{m_1} \otimes \alpha \otimes \cdots \otimes \alpha \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & \sum_{N=\hat{K}(\mathbf{k})}^{\infty} p(N,\mathbf{k},\beta) D_{1+K+N} \otimes \mathbf{I}_{m_1^{s-1}} \otimes \alpha \\ 0 & \dots & 0 & \sum_{N=\hat{K}(\mathbf{k})}^{\infty} p(N,\mathbf{k},\beta) D_{K+N} \otimes \mathbf{I}_{m_1^{s}} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & \sum_{N=\hat{K}(\mathbf{k})}^{\infty} p(N,\mathbf{k},\beta) D_{K+N} \otimes \mathbf{I}_{m_1^{s}} \end{pmatrix}$$

$$\equiv \sum_{N=\hat{K}(\mathbf{k})}^{\infty} \big[p(N,\mathbf{k},\beta) A_0(N) \big], \tag{6.1}$$

where

$$\widehat{K}(\mathbf{k}) = \sum_{j=1}^{m_2} k_j,$$

$$p(N,\mathbf{k},\beta) = \frac{N!}{(N-\widehat{K}(\mathbf{k}))! k_1! \cdots k_{m_2}!} p^{\hat{K}(\mathbf{k})} (1-p)^{N-\hat{K}(\mathbf{k})} \beta_1^{k_1} \cdots \beta_{m_2}^{k_{m_2}},$$

$$B_{ij} = \begin{cases} \displaystyle\sum_{N=j-i}^{\infty} \binom{N}{j-i} p^{j-i} (1-p)^{N-j+i} (D_N \otimes \mathbf{I}_{m_1^i}) \otimes \underbrace{\alpha \otimes \cdots \otimes \alpha}_{j-i}, \\ \quad 0 \leqslant i < j \leqslant s, \\ \displaystyle\sum_{N=j-i}^{\infty} \binom{N}{j-i} p^{j-i} (1-p)^{N-j+i} (D_N \otimes \mathbf{I}_{m_1^i}) \otimes \underbrace{\alpha \otimes \cdots \otimes \alpha}_{s-i}, \\ \quad 0 \leqslant i < s \leqslant j \leqslant s+K, \\ \displaystyle\sum_{N=j-i}^{\infty} \binom{N}{j-i} p^{j-i} (1-p)^{N-j+i} (D_N \otimes \mathbf{I}_{m_1^s}), \\ \quad s \leqslant i < j \leqslant s+K, \end{cases} \tag{6.2}$$

and

$$\binom{N}{x} = \frac{N!}{x!(N-x)!}.$$

The necessity of the condition for ergodicity can be proved by comparing the retrial queue with the BMAP/PH/s queue with customer loss at arrival epochs, i.e., every customer leaves the system upon arrival with probability p. The sample path method used in section 4 can be used again to prove the result. Details are omitted.

To prove the sufficiency, use the same test function defined in inequality (5.3). Inequality (5.8) then becomes

$$\begin{aligned}
&z^{\sum_{i=1}^{m_2} n_i - 1}\Bigg\{ a\sum_{N=1}^{\infty} z\big[(zp+1-p)^N - 1\big]A_0(N)\mathbf{e} - a(z-1)\left(\sum_{i=1}^{m_2} n_i h_i^0\right)\Gamma\mathbf{e} \\
&\quad + z\Bigg[A_1 + \sum_{N=1}(zp+1-p)^N A_0(N)\Bigg]\mathbf{u} + \left(\sum_{i=1}^{m_2} n_i h_i^0\right)(A_2 - z\Gamma)\mathbf{u}\Bigg\} \\
&\quad + a(z-1)\sum_{i\in\Im(1)} n_i\left[\sum_{t=2}^{j_0} z^{\sum_{l\in\Re(t)} n_l}\left(\sum_{l\in\Re(t)} h_{i,l}\right)\right]\mathbf{e} \\
&\quad + a\sum_{j=2}^{j_0}\Bigg\{ z^{\sum_{t\in\Re(j)} n_t}\sum_{\mathbf{k}:\,\exists k_l>0,\ l\in\Re(j)}\left(z^{\sum_{t\in\Re(j)} k_t} - 1\right)\left[\sum_{N=\bar{K}(\mathbf{k})}^{\infty} p(N,\mathbf{k},\beta)A_0(N)\right]\mathbf{e} \\
&\quad + \sum_{i\in\Im(j)} n_i\left[\sum_{l=j+1}^{j_0} z^{\sum_{t\in\Re(j)} n_t}\left(\sum_{t\in\Re(l)} h_{i,t}\right) - z^{\sum_{t\in\Re(j)} n_t - 1}\sum_{t\in\Im(j-1)} h_{i,t}\right]\mathbf{e}\Bigg\} \leqslant -\varepsilon\mathbf{e}.
\end{aligned} \tag{6.3}$$

Again, consider the BMAP/PH/s queue in which every customer leaves the system upon arrival with probability p. When $\rho = (p\lambda)/(s\mu) < 1$, this queueing system is ergodic. The rest of the proof is similar to that in section 5. Details are omitted. This completes the proof. □

Retrial queues with customer loss at both arrival epochs and retrial epochs. Consider the BMAP/PH/$s/s + K$ retrial queueing system with PH-retrial times and customer loss at arrival epochs. Assume that when a retrial customer finds no server and no waiting position available, the customer reenters the orbit with probability q and leaves the queueing system with probability $1 - q$, $0 \leqslant q \leqslant 1$.

Theorem 5. The BMAP/PH/$s/s + K$ retrial queueing system with PH-retrial times, customer loss at both arrival epochs and retrial epochs, is ergodic if and only if either $q < 1$ or $q = 1$ and $\rho = (p\lambda)/(s\mu) < 1$.

Proof. When $q = 1$, the theorem reduces to theorem 4. When $q < 1$, the necessity is clear from theorem 4. To prove sufficiency, consider the Markov process $\{N_i(t), 1 \leqslant i \leqslant m_2, q(t), I(t), I_j(t), 1 \leqslant j \leqslant \min\{s, q(t)\}\}$ defined in section 3. The infinitesimal generator of this Markov process is the same as that in theorem 4 except

$$\begin{aligned} A(\mathbf{n}, \mathbf{n}) &= A_1 + \sum_{i=1}^{m_2} n_i h_{i,i} \mathbf{I} + \sum_{i=1}^{m_2} n_i h_i^0 q(\mathbf{I} - \Gamma), \\ A(\mathbf{n}, \mathbf{n} - \mathbf{e}_i) &= n_i h_i^0 A_2 + n_i h_i^0 (1 - q)(\mathbf{I} - \Gamma). \end{aligned} \tag{6.4}$$

Using the test functions defined in inequality (5.3), inequality (5.8) becomes

$$\begin{aligned} & z^{\sum_{i=1}^{m_2} n_i - 1} \Bigg\{ a \sum_{N=1}^{\infty} z \big[(zp + 1 - p)^N - 1 \big] A_0(N) \mathbf{e} \\ & \quad - a(z-1) \left(\sum_{i=1}^{m_2} n_i h_i^0 \right) \big[\Gamma + (1-q)(\mathbf{I} - \Gamma) \big] \mathbf{e} \\ & \quad + z \left[A_1 + \sum_{N=1} (zp + 1 - p)^N A_0(N) \right] \mathbf{u} \\ & \quad + \left(\sum_{i=1}^{m_2} n_i h_i^0 \right) \big[A_2 + (1-z)(1-q)(\mathbf{I} - \Gamma) - z\Gamma \big] \mathbf{u} \Bigg\} \\ & \quad + a \sum_{i \in \Im(1)} n_i \left[\sum_{l=2}^{m_2} z^{\sum_{t \in \Re(l)} n_t} \left(\sum_{t \in \Re(l)} h_{i,t} \right) \right] \mathbf{e} \\ & \quad + a \sum_{j=2}^{j_0} z^{\sum_{l \in \Re(j)} n_l} \left\{ \sum_{\mathbf{k}:\, \exists k_l > 0,\ l \in \Re(j)} \left(z^{\sum_{l \in \Re(j)} k_l} - 1 \right) \left[\sum_{N = \hat{K}(\mathbf{k})}^{\infty} p(N, \mathbf{k}, \beta) A_0(N) \right] \mathbf{e} \right. \\ & \quad \left. + \sum_{i \in \Im(j)} n_i \left[\sum_{l=j+1}^{m_2} z^{\sum_{t \in \Re(j)} n_t} \left(\sum_{t \in \Re(l)} h_{i,t} \right) - z^{\sum_{t \in \Re(j)} n_t - 1} \sum_{t \in \Im(j-1)} h_{i,t} \right] \mathbf{e} \right\} \leqslant -\varepsilon \mathbf{e}. \end{aligned} \tag{6.5}$$

It is clear that when $q < 1$, a set of parameters can always be found so that equation (6.5) holds for all but a finite number of $\mathbf{n} \in \mathcal{N}$ for some positive ε. For instance, choose z to be close to one (and $z > 1$) and $\mathbf{u} = 0$. Other parameters can be determined accordingly. Notice that the test function of this case can be simpler, since vectors $\{\mathbf{f}_{\mathbf{n}}(2), \mathbf{n} \in \mathcal{N}\}$ do not have to be included in the test function. This completes the proof. □

It is interesting to see that such a queueing model is always ergodic when $q < 1$. Intuitively, since customers can be lost upon retrials, on average, more customers will be lost per unit time when more customers are in the orbit. Therefore, the number of

customers in the orbit will not go to infinity. Then the Markov process is ergodic and so the queueing system.

7. Ergodicity of approximation models

This section studies the ergodicity of two modified queueing systems which were introduced as approximations to the stationary distribution of the queueing system of interest (see [5,6,9,21]). One of the two modified queues has a smaller retrial rate than the original retrial queueing system, while the other queue has a larger retrial rate.

The lower-bound queue. For a fixed nonnegative integer N, define a retrial queue similar to the BMAP/PH/s/s+K retrial queue with PH-retrial times except that when there are more than N customers in the orbit retrials become instant. That is, when there are more than N customers in the orbit and a service is complete, a customer in the orbit is selected to enter the queue (or the server) immediately. The selection of such a customer can be arbitrary (an issue that does not need to be addressed explicitly in the proof of theorem 6). This queue is called a lower-bound queue since it is *believed* that it has a smaller total number of customers in service, the queue, or the orbit.

Theorem 6. For any nonnegative integer N, The lower-bound queue is ergodic if and only if $\rho < 1$.

Proof. The necessity of $\rho < 1$ can be proved by the sample path method used in theorem 2. The sufficiency of $\rho < 1$ can be shown by coupling the lower-bound queue with the BMAP/PH/s queueing system when the total number of customers in the system is larger than N. When the total number of customers in the lower-bound queue is larger than N, the queueing process reduces to that of the BMAP/PH/s queueing system. Details are omitted. This completes the proof. □

The upper-bound queue. For a fixed positive integer N, define a retrial queueing system similar to the BMAP/PH/s/s + K retrial queue with PH-retrial times except that at most N customers in the orbit are allowed to retry for service at any time. If a customer, who must enter the orbit, finds that there are N customers in the orbit, the customer does not enter the orbit until the number of customers in the orbit becomes less than N. For customers waiting to enter the orbit, getting into the orbit follows a first-in-first-entering rule. This queue is called an upper-bound queue since it is *believed* that it has a larger total number of customers in service, the queue, waiting to enter the orbit, or the orbit.

Theorem 7. When N is large enough, the upper-bound queue is ergodic if and only if $\rho < 1$.

Proof. The necessity of $\rho < 1$ can be proved by the sample path method used in theorem 2. To prove sufficiency, consider the Markov process $\{N_i(t),\ 0 \leqslant i \leqslant m_2,\ q(t),\ I(t),\ I_j(t),\ 1 \leqslant j \leqslant \min\{s, q(t)\}\}$ introduced in section 3 except that $N_0(t)$ is introduced to record the number of customers who are waiting to enter the orbit and $N_1(t)+\cdots+N_{m_2}(t) \leqslant N$. $N_0(t)$ can be positive only when $N_1(t)+\cdots+N_{m_2}(t) = N$. In the proof to theorem 3, when one value in $\{n_i\}$ ($\mathbf{n} = (n_1, \ldots, n_{m_2})$) is large enough, inequality (5.2) holds. Suppose that inequality (5.2) holds for $n_i > n_i^*$ for each i ($1 \leqslant i \leqslant m_2$). Choose N that is larger than the sum of $\{n_i^* + 1\}$. Then the mean-drift method can be applied to the corresponding Markov process $\{N_i(t),\ 0 \leqslant i \leqslant m_2,\ q(t),\ I(t),\ I_j(t),\ 1 \leqslant j \leqslant \min\{s, q(t)\}\}$ and inequality (5.2) holds for all but a finite states. In fact, according to the proof to theorem 3, inequality (5.2) holds whenever $N_0(t) > 0$, since $N_0(t) > 0$ implies that $N_1(t) + \cdots + N_{m_2}(t) = N$ so that at least one of $\{N_1(t), \ldots, N_{m_2}(t)\}$ is larger than its corresponding n_i^*. Details are omitted. This completes the proof. □

Note. When $N = 1$, the upper-bound queue is also known as a retrial queue in which only the customer at the head of the retrial line (queue) can retry for service. Studies of this type of retrial queues can be found in [3] (also see [1,12]). When $N = 1$, the Markov process becomes a quasi-birth-and-death Markov process and its ergodicity condition can be found using Neuts' condition. Details are omitted. It is interesting to find the ergodicity condition for the upper bound retrial queue when N is not large.

8. Some discussion

Theorems 1, 6, and 7 show that the BMAP/PH/s/$s + K$ retrial queue with PH-retrial times, the two modifications (lower-bound and upper-bound queues), and the classical BMAP/PH/s queue are ergodic if and only if $\rho < 1$. Why is the condition $\rho < 1$ a necessary and sufficient condition for ergodicity of the four quite different queueing systems? We offer some intuition to this question. Notice that "retrial" delays the service of a customer in the orbit. One of the consequences is that idle periods (of servers) are different for the four queueing systems. For retrial queues, a server may become idle frequently for a period of time when a small number of customers are in the orbit. On the other hand, the server may be busy for a long time or its idle times are cut short when a lot of customers are in the orbit. On average, the ratio of the total idle time to the total busy time of a server remains the same for the four queueing systems. In a retrial queue, a server may be idle while there are customers in the orbit trying for service. Thus, a retrial queueing system may lose some service capacity when the number of customers in the orbit is not large. Fortunately, the loss of capacity is recovered when the number of customers in the orbit becomes large. In this case, servers of the retrial queues have to serve customers from outside as well as retrial customers who seize any idle server almost instantly.

A special case – the MAP/PH/s/s+K retrial queue with exponential retrial times ($m_2 = 1$) – is of special importance because (1) the Markov process $\{N_1(t),\ q(t),\ I(t),$

$I_j(t)$, $1 \leqslant j \leqslant \min\{s, q(t)\}\}$ is a quasi-birth-and-death Markov process, and (2) it has the M/M/s/s retrial queue with exponential retrial times, and the MAP/PH/1/1 retrial queue with exponential retrial times as its special cases. Its corresponding lower-bound and upper-bound retrial queues have matrix–geometric solutions. Theorems 6 and 7 present the condition to ensure the existence of the matrix–geometric solutions.

Acknowledgements

The authors would like to thank referees for their valuable comments and suggestions. This work has been financially supported through NSERC (National Science and Engineering Research Council of Canada) operating grants of the three authors.

References

[1] J.R. Artalejo, Accessible bibliography on retrial queues, Math. Comput. Modelling 30 (1999) 1–6.
[2] S. Asmussen, *Applied Probability and Queues* (Wiley, Chichester/New York, 1987).
[3] B.D. Choi, K.K. Park and C.E.M. Pearce, An M/M/1 retrial queue with control policy and general retrial times, Queueing Systems 14 (1993) 275–292.
[4] J.W. Cohen, *The Single Server Queues* (North-Holland, Amsterdam, 1982).
[5] J. Diamond, Matrix analytic methods for retrial queues, Ph.D. thesis, Department of Mechanical and Industrial Engineering, University of Manitoba (1995).
[6] J. Diamond and A.S. Alfa, Matrix analytic methods for M/PH/1 retrial queues, Stochastic Models 11 (1995) 447–470.
[7] J. Diamond and A.S. Alfa, The MAP/PH/1 retrial queue, Stochastic Models 14 (1998) 1151–1178.
[8] J.E. Diamond and A.S. Alfa, Matrix analytical methods for retrial queues with finite buffers, submitted to Queueing Systems.
[9] J.E. Diamond and A.S. Alfa, Matrix analytical method for multiserver retrial queues with finite buffers, accepted *Workshop on Retrial Queues (WRQ'98)*, Madrid (September 1998).
[10] G.I. Falin, On sufficient conditions for ergodicity of multi-channel queueing systems with repeated calls, Adv. in Appl. Probab. 16 (1984) 447–448.
[11] G.I. Falin, A survey of retrial queues, Queueing Systems 7 (1990) 127–167.
[12] G.I. Falin and J.G.C. Templeton, *Retrial Queues* (Chapman & Hall, London, 1997).
[13] F.R. Gantmacher, *The Theory of Matrices* (New York, Chelsea, 1959).
[14] V.G. Kulkarni and H.M. Liang, Retrial queues revisited, in: *Frontiers in Queueing: Models and Applications in Science and Engineering,* ed. J.H. Dshalalow (CRC Press, Boca Raton, FL, 1997) pp. 19–34.
[15] H. Li and T. Yang, The steady-state distribution of the PH/M/1 retrial queue, in: *Advances in Matrix Analytic Methods for Stochastic Models*, eds. A.S. Alfa and S.R. Chakravarthy (Notable Publications, 1998) pp. 135–150.
[16] H.M. Liang and V.G. Kulkarni, Stability condition for a single server retrial queue, Adv. in Appl. Probab. 25 (1993) 690–701.
[17] D.M. Lucantoni, New results on the single server queue with a batch Markovian arrival process, Stochastic Models 7 (1991) 1–46.
[18] M.F. Neuts, A versatile Markovian point process, J. Appl. Probab. 16 (1979) 764–779.
[19] M.F. Neuts, *Matrix–Geometric Solutions in Stochastic Models: An Algorithmic Approach* (Johns Hopkins Univ. Press, Baltimore, MD, 1981).
[20] M.F. Neuts, *Structured Stochastic Matrices of* M/G/1 *Type and Their Applications* (Marcel Dekker, New York, 1989).

[21] M.F. Neuts and B.M. Rao, Numerical investigation of a multiserver retrial model, Queueing Systems 7 (1990) 169–189.
[22] T. Yang, M.J.M. Posner, J.G.C. Templeton and H. Li, An approximation method for the M/G/1 retrial queue with general retrial times, European J. Oper. Res. 76 (1994) 552–562.
[23] T. Yang and J.G.C. Templeton, A survey on retrial queues, Queueing Systems 2 (1987) 201–233.