Ergodicity of the *BMAP*/*PH*/*s*/*s*+*K* Retrial Queue with *PH*-Retrial Times

[1]Qi-Ming HE, [2]Hui Li, and [3]Yiqiang Q. Zhao

[1]Department of Industrial Engineering, Dalhousie University
Halifax, Nova Scotia Canada B3J 2X4

[2]Department of Mathematics, Mount Saint Vincent University
Halifax, Nova Scotia Canada B3M 2J6

[3]Department of Mathematics and Statistics, University of Winnipeg
Winnipeg, Manitoba Canada R3B 2E9

(June, 1998)

AMS 1991 Subject Classification: 60K25

Abstract 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.

Key words: Retrial queue, ergodicity, sample path method, mean-drift method, matrix analytic methods, batch Markov arrival process (*BMAP*), *PH*-distribution, and impatient customer, Lyapunov function.

1. Introduction

The retrial queueing system studied in this paper has finite waiting positions and a number of servers. When an arriving customer finds that all servers are busy and no waiting position is available, the customer starts orbiting in 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 study the ergodicity of such retrial queueing systems.

Ergodicity of retrial queues has been studied by many researchers (see Falin [9], Falin and Templeton [10], Kulkarni and Liang [12], Yang and Templeton [20], and references therein). Conditions 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 Falin [8], the sufficiency of $\rho<1$ for ergodicity of the *M/M/s/s* retrial queue with exponential retrial times was proved. Liang and Kulkarni [13] obtained a stability condition of a single server retrial queue. Yang, et al. [21] and Diamond [3] showed that $\rho<1$ is a necessary and sufficient condition for ergodicity of the *M/G/1* retrial queue with general retrial times, respectively. For single server retrial queues with a Markov arrival process, *PH*-service times, and exponential retrial times or their special cases, Diamond [3], Diamond and Alfa [4] and [5], and Li and Yang [14] proved that $\rho<1$ is necessary and sufficient for ergodicity. Diamond and Alfa [6] and [7] proved that $\rho<1$ is a sufficient condition for ergodicity of multi-server retrial queues with finite buffers and exponential retrial times (also see Diamond [3]). These results support a simple and intuitive conjecture on ergodicity of retrial queues which has been used by many researchers without rigorous proof.

Conjecture 1.1 A retrial queueing system is ergodic if and only if $\rho<1$.

By that a queueing system is ergodic, we mean that an associated Markov process (defined later) of the queueing system is ergodic. It has been found that condition $\rho<1$ is not enough for the ergodicity of some retrial queues, one of which is shown below (see Liang and Kulkarni [13] for more counterexamples).

Example 1.2 Consider a single server retrial queueing system with deterministic interarrival times and deterministic retrial times, both of which have the same length one. The service time is *0.1* with probability *0.9* and *5.1* with probability *0.1*. There is no waiting position. 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 go orbiting. On the other hand, no orbiting customer 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.

Despite of these counterexamples, it is still believed that Conjecture 1.1 is true for retrial queues whose interarrival times, service times, and retrial times have continuous distribution functions. 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 arrival process, *PH*-service times, multiple servers, finite waiting positions, and *PH*-retrial times, which can be considered as a generalization of the retrial queues studied in [3], [4], [5], [6], [7], [8], [12], and [19]. Among them, the *MAP/PH/s/s+K* retrial queue with exponential retrial times considered in Diamond and Alfa [6] and [7] is the one closest to the model studied 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 [10] and also see Foster's criterion in Cohen [2]) is used to prove the sufficiency of the condition. 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 in fact challenging. 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 queueing system of interest is modelled into a highly structured Markov process, which makes it possible to prove the sufficiency of the condition for ergodicity. This Markov process can also be used to study the stationary distribution of the retrial queueing systems of interest.

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. Finally, 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 $\rho<1$.

2. The *BMAP/PH/s/s+K* Retrial Queue with *PH*-Retrial Times

The basic 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 to the queueing system according to a batch Markov arrival process. The batch Markov arrival process (*BMAP*) was introduced by Neuts (see Neuts [16] and [18] and Lucantoni [15]) as a generalization of the phase-type renewal process (see Neuts [17]). 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 $(-D_0)_{i,i}$ $(\geq -(D)_{i,i})$. 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_0 be the rate matrix of transitions that does not generate arrivals, D_1 the rate matrix of transitions with one customer, D_2 the rate matrix of transitions with two customers, etc. Notice that the matrix D_0 has strictly negative diagonal elements and nonnegative off-diagonal elements, matrices $\{D_n, n>0\}$ are nonnegative, and $D = D_0 + \Sigma_{n\geq 1} D_n$. 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 $D^*(z) = \sum_{n=0}^{\infty} z^n D_n$. Assume that $D^*(z)$ is finite for $0<z<z_0$, where $z_0 > 1$. This condition is not restrictive since $D^*(z)$ is always finite when only a finite number of matrices in $\{D_n, n\geq 1\}$ are nonzero.

There are s identical servers serving customers one 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 vector of size m_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 Chapter 2 of Neuts [17]. 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 vector of size m_2, and $H = (h_{i,j})$ is an $m_2 \times m_2$ matrix. Define $\mathbf{H}^0 = -H\mathbf{e} = (h_i^0)$. 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. Any orbiting customer must be in one of the m_2 states of the *PH*-distribution at any time.

3. The Infinitesimal Generator

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

- $N_i(t)$ be the number of customers in the orbit whose retrial time process is in state i at time t, $1 \le i \le m_2$;
- $q(t)$ 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 server which is working at time t, $1 \le j \le \max\{s, q(t)\}$.

It is easy to see that $\{N_i(t), 1 \le i \le m_2, q(t), I(t), I_j(t), 1 \le j \le \max\{s, q(t)\}\}$ is an irreducible Markov process. Let $\aleph = \{\mathbf{n} = (n_1, \ldots, n\_\{m_2\}): 0 \le n_i < \infty, 1 \le i \le m_2\}$. Notation "*x_y*" is used for *x* subscript *y* for typographical reasons. The state space of the Markov process is

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

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

For convenience, transitions of the Markov process $\{N_i(t), 1\le i\le m_2, q(t), I(t), I_j(t), 1 \le j \le \max\{s, q(t)\}\}$ 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, $1\le i\le m_2$, 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 $1\le i, j\le m_2$, in one transition. The infinitesimal generator of Markov process $\{N_i(t), 1\le i\le m_2, q(t), I(t), I_j(t), 1 \le j \le \max\{s, q(t)\}\}$ 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

$$A(\mathbf{n},\mathbf{n}+\mathbf{k}) = \begin{pmatrix} 0 & \cdots & 0 & p(\mathbf{k},\boldsymbol{\beta})D_{s+K+\sum_{j=1}^{m_2}k_j} \otimes \underbrace{\boldsymbol{\alpha}\otimes\cdots\otimes\boldsymbol{\alpha}}_{s} \\ 0 & \cdots & 0 & p(\mathbf{k},\boldsymbol{\beta})D_{s-1+K+\sum_{j=1}^{m_2}k_j} \otimes \mathbf{I}_{m_1} \otimes \underbrace{\boldsymbol{\alpha}\otimes\cdots\otimes\boldsymbol{\alpha}}_{s-1} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & p(\mathbf{k},\boldsymbol{\beta})D_{1+K+\sum_{j=1}^{m_2}k_j} \otimes \mathbf{I}_{m_1^{s-1}} \otimes \boldsymbol{\alpha} \\ 0 & \cdots & 0 & p(\mathbf{k},\boldsymbol{\beta})D_{K+\sum_{j=1}^{m_2}k_j} \otimes \mathbf{I}_{m_1^{s}} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & p(\mathbf{k},\boldsymbol{\beta})D_{\sum_{j=1}^{m_2}k_j} \otimes \mathbf{I}_{m_1^{s}} \end{pmatrix} \tag{3.2}$$

$$\equiv p(\mathbf{k},\boldsymbol{\beta})A_0\left(\sum_{j=1}^{m_2}k_j\right),$$

where

$$p(\mathbf{k},\boldsymbol{\beta}) = \frac{(k_1+\cdots+k_{m_2})!}{k_1!\cdots k_{m_2}!}\beta_i^{k_1}\cdots\beta_{m_2}^{k_{m_2}}, \tag{3.3}$$

$x! = x(x-1)\cdots 1$, $0! = 1$, and "$\otimes$" denotes the Kronecker product (Gantmacher [11]). $A(\mathbf{n}, \mathbf{n}+\mathbf{k})$ is an $M\times M$ matrix. Note that the blocks within $A(\mathbf{n}, \mathbf{n}+\mathbf{k})$ describe the transitions 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 go orbiting. For customers who go orbiting, the selection of the initial states of their retrial times follows the multinomial distribution $\{p(\mathbf{k}, \boldsymbol{\beta})\}$.

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

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

where Γ is an $M\times M$ matrix with all diagonal elements to be one except the last mm_1^s elements, which are zero, and all off-diagonal elements 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

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

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

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

Note that $[\Sigma_{n\geq 1}A_0(n)+A_1]\mathbf{e} = 0$, 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 ($j\leq s+K$), 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 $1\leq i\leq m_2$ and $n_i>0$,

$$
\begin{aligned}
&A(\mathbf{n},\mathbf{n}-\mathbf{e}_i)\\
&= n_i h_i^0 \begin{pmatrix}
0 & \mathbf{I}_m \otimes \alpha & & & & & & & \\
 & 0 & \mathbf{I}\_\{mm_1\} \otimes \alpha & & & & & & \\
 & & \ddots & \ddots & & & & & \\
 & & & 0 & \mathbf{I}\_\{mm_1^{s-1}\} \otimes \alpha & & & & \\
 & & & & 0 & \mathbf{I}\_\{mm_1^s\} & & & \\
 & & & & & \ddots & \ddots & & \\
 & & & & & & 0 & \mathbf{I}\_\{mm_1^s\} & \\
 & & & & & & & 0 &
\end{pmatrix} \qquad (3.8)\\
&\equiv n_i h_i^0 A_2.
\end{aligned}
$$

Note that $A(\mathbf{n}, \mathbf{n}\text{-}\mathbf{e}_i)$ is an $M \times M$ matrix and $(A_2\text{-}\Gamma)\mathbf{e} = 0$. Also notice that a retrial is successful only when the total number of customer in queue or service is less than $s+K$. Although the infinitesimal generator is complicated, its construction is straightforward and explicit. Thus, no more explanations are given to this construction process.

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 is ergodic if and only if $\rho<1$.

The proof of Theorem 1 consists of two parts: 1) a proof 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 $\rho<1$ for ergodicity of the queueing system of interest, 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 3.

Theorem 2. If the *BMAP*/*PH*/*s*/*s*+*K* retrial queue with *PH*-retrial times 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 lemma. Suppose that the retrial queue and the corresponding non-retrial queue *BMAP/PH/s* are empty initially. Let these two queueing processes be coupled in the same probability space. Let a_n be the arrival epoch of the *n*th customer and s_n the service time of the *n*th service. Notice that s_n may not be the *n*th customer's service time in the retrial queue. Let t_n and $t_{L,n}$ be the epochs when the nth service starts for the retrial queue and the corresponding non-retrial queue, respectively. It is clear that, for $n>s$,

$$\begin{aligned} t_{L,n} &= \max\{a_n, \min\{t_{L,n-s}+s_{n-s}, t_{L,n-s+1}+s_{n-s+1}, \cdots, t_{L,n-1}+s_{n-1}\}\}; \\ t_n &\geq \max\{a_n, \min\{t_{n-s}+s_{n-s}, t_{n-s+1}+s_{n-s+1}, \cdots, t_{n-1}+s_{n-1}\}\}. \end{aligned} \quad (4.1)$$

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

$$\begin{aligned} &A(t) = \max\{n:\ a_n < t\}; \\ &B(t) = \max\{n:\ t_n + s_n < t\}, \quad B_L(t) = \max\{n:\ t_{L,n} + s_n < t\}; \\ &q_a(t) = A(t) - B(t), \qquad\qquad q_{L,a}(t) = A(t) - B_L(t). \end{aligned} \quad (4.2)$$

It is easy to see that $B(t) \leq B_{L,}(t)$ and therefore $q_{all}(t) \geq q_{L,all}(t)$ for all t. This implies that the total number of customers in the $MAP/PH/s/s+K$ retrial queue is always as large as that in the $MAP/PH/s$ queue. This further implies that $\mathbf{P}\{q_{all}(t) \leq q\} \leq \mathbf{P}\{q_{L,all}(t) \leq q\}$ for all $q \geq 0$ and $t>0$. Setting $q=0$ yields $\mathbf{P}\{q_{all}(t) = 0\} \leq \mathbf{P}\{q_{L,all}(t) = 0\}$. When the retrial queue is ergodic, the limit $\lim_{\{t\to\infty\}}\mathbf{P}\{q_{all}(t) = 0\}$ exists and is positive. Since the Markov process $\{q_{L,all}(t), I(t), I_i(t), 1 \leq i \leq \max\{s, q_{L,all}(t)\}\}$ of the $MAP/PH/s$ queue is irreducible, the limit $\lim_{\{t\to\infty\}}\mathbf{P}\{q_{L,all}(t)=0\}$ exists and is positive since $\lim_{\{t\to\infty\}}\mathbf{P}\{q_{L,all}(t)=0\} \geq \lim_{\{t\to\infty\}}\mathbf{P}\{q_L(t)=0\}>0$. This implies that the $MAP/PH/s$ queue is ergodic. When the $MAP/PH/s$ queue is ergodic, $\rho<1$ must be true (Asmussen [1]). 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, the mean-drift method (or Foster's criterion) is utilized (see Falin and Templeton [10] and Cohen [1]). The result is stated in the following theorem and its proof is rather long.

Theorem 3. When $\rho<1$, the Markov process $\{N_i(t), 1 \leq i \leq m_2, q(t), I(t), I_j(t), 1 \leq j \leq \max\{s, q(t)\}\}$ introduced in Second 3 is ergodic.

Proof. To prove the theorem by using the mean-drift method, the key is to construct a vector-valued test (or Lyapunov) function $\{\mathbf{f_n}, \mathbf{n} \in \aleph\}$ such that $\mathbf{f_n} \to \infty$ when $\sum_{i=1}^{m\_2} n_i \to \infty$ and

$$\sum_{i=1}^{m\_2} A(\mathbf{n},\mathbf{n}-\mathbf{e}_i)\mathbf{f}_{\mathbf{n}-\mathbf{e}\_i} + A(\mathbf{n},\mathbf{n})\mathbf{f}_{\mathbf{n}} + \sum_{i=1}^{m\_2}\sum_{j=1:\, j\neq i}^{m\_2} A(\mathbf{n},\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j)\mathbf{f}_{\mathbf{n}-\mathbf{e}\_i+\mathbf{e}\_j} + \sum_{\mathbf{k}\geq 0,\, \mathbf{k}\neq 0} A(\mathbf{n},\mathbf{n}+\mathbf{k})\mathbf{f}_{\mathbf{n}+\mathbf{k}} \leq -\varepsilon\mathbf{e} \tag{5.1}$$

holds for all but a finite number of $\mathbf{n}\in\aleph$ for some positive ε. In the above inequality and the following, $A(\mathbf{n}, \mathbf{n}') = 0$ if $\mathbf{n}'\notin\aleph$. According to the mean-drift theory of Markov processes, if inequality (5.1) holds for all but a finite number of $\mathbf{n}\in\aleph$ for some positive ε, the corresponding Markov process is ergodic. For such a purpose, the following test function is introduced. Notation "x^y" shall be used for x superscript y for typographical reasons. For $\mathbf{n}\in\aleph$,

$$\mathbf{f}_{\mathbf{n}} = a\sum_{j=1}^{j\_0} z\wedge\{\sum_{i\in\Re(j)} n_i\}\mathbf{e} + z\wedge\{\sum_{i=1}^{m\_2} n_i\}\mathbf{u} \equiv \sum_{j=1}^{j\_0}\mathbf{f}_{\mathbf{n}}(1, j) + \mathbf{f}_{\mathbf{n}}(2), \tag{5.2}$$

where $1<z<z_0$, $\mathbf{u}$ is a vector of size M, a is a positive number, j_0 is a positive integer, $\Im(1) = \{i: 1\leq i\leq m_2,\ h_i^0 \neq 0\}$ and, for $2\leq j\leq j_0$,

$$\begin{aligned}&\Im(j) = \{i:\ 1\leq i\leq m_2,\ i\notin\bigcup_{k=1}^{j-1}\Im(k),\ \text{and for some } t\in\Im(j-1),\ h_{i,t}>0\};\\ &\Re(j) = \bigcup_{k=j}^{j\_0}\Im(k),\quad \Re(1)-\Re(j)\equiv\bigcup_{k=1}^{j-1}\Im(k),\quad \text{and}\quad \Re(1)=\{1, 2, ..., m_2\}.\end{aligned} \tag{5.3}$$

If $\Im(1)=\{1, 2, \ldots, m_2\}$, then $\Im(1)=\Re(1)$ and all other sets are empty. Intuitively, 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 $1\leq j\leq j_0$. In general, the retrial process of a customer can go from a state in $\Im(j)$ to a state in $\Im(j-1)$, $\Im(j)$, …, $\Im(j_0-1)$, or $\Im(j_0)$, but not any state in $\Im(1)$, …, or $\Im(j-2)$ after one transition. Subsets $\{\Im(j), 1\leq j\leq j_0\}$ play an important role in the following proof. Understanding the transitions of the retrial process among these subsets shall be helpful. Notice that when the (arrival) batch size is one and retrial times are exponential, Diamond and Alfa [7] introduced a test function similar to the one given in equation (5.2) with $j_0=1$, $m_2=1$, and $\Im(1)=\Re(1)=\{1\}$ and proved this theorem.

Values of parameters z, $\mathbf{u}$, and a shall be determined so that inequality (5.1) holds for all but a finite number of $\mathbf{n}\in\aleph$ for some positive ε. For this purpose, the left hand side of inequality (5.1) is evaluated as follows. First, terms containing vectors $\{\mathbf{f}_{\mathbf{n}}(1,1)\}$ on the left hand side of inequality (5.1) are evaluated.

$$\sum_{i=1}^{m\_2} A(\mathbf{n},\mathbf{n}-\mathbf{e}_i)\mathbf{f}_{\mathbf{n}-\mathbf{e}\_i}(1,1) + A(\mathbf{n},\mathbf{n})\mathbf{f}_{\mathbf{n}}(1,1) + \sum_{i=1}^{m\_2}\sum_{j=1:\, j\neq i}^{m\_2} A(\mathbf{n},\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j)\mathbf{f}_{\mathbf{n}-\mathbf{e}\_i+\mathbf{e}\_j}(1,1) + \sum_{\mathbf{k}\geq 0,\mathbf{k}\neq 0} A(\mathbf{n},\mathbf{n}+\mathbf{k})\mathbf{f}_{\mathbf{n}+\mathbf{k}}(1,1) \tag{5.4}$$

$$= z^{\wedge}\{\sum_{i=1}^{m\_2} n_i - 1\}\left[\sum_{i=1}^{m\_2} A(\mathbf{n},\mathbf{n}-\mathbf{e}_i) + zA(\mathbf{n},\mathbf{n}) + z\sum_{i=1}^{m\_2}\sum_{j=1:\ j\neq i}^{m\_2} A(\mathbf{n},\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j) + \sum_{\mathbf{k}\geq 0,\,\mathbf{k}\neq 0} z^{\wedge}\{1+\sum_{i=1}^{m\_2} k_i\} A(\mathbf{n},\mathbf{n}+\mathbf{k})\right] a\mathbf{e}$$

$$= z^{\wedge}\{\sum_{i=1}^{m\_2} n_i - 1\}\left[(\sum_{i=1}^{m\_2} n_i h_i^0) A_2 + zA_1 + z\sum_{i=1}^{m\_2} n_i h_{i,i}\mathbf{I} + z\sum_{i=1}^{m\_2} n_i h_i^0(\mathbf{I}-\Gamma) + z\sum_{i=1}^{m\_2}\sum_{j=1:\ j\neq i}^{m\_2} n_i h_{i,j}\mathbf{I} + \sum_{\mathbf{k}\geq 0,\,\mathbf{k}\neq 0} z^{\wedge}\{1+\sum_{i=1}^{m\_2} k_i\} p(\mathbf{k},\beta) A_0(\sum_{j=1}^{m\_2} k_j)\right] a\mathbf{e}$$

$$= z^{\wedge}\{\sum_{i=1}^{m\_2} n_i - 1\}\left\{(\sum_{i=1}^{m\_2} n_i h_i^0) A_2 + zA_1 + z\sum_{i=1}^{m\_2} n_i h_{i,i}\mathbf{I} + z\sum_{i=1}^{m\_2} n_i h_i^0(\mathbf{I}-\Gamma) + z\sum_{i=1}^{m\_2}\sum_{j=1:\ j\neq i}^{m\_2} n_i h_{i,j}\mathbf{I} + \sum_{N=1}^{\infty} z^{1+N}[(\sum_{\mathbf{k}:\ \sum_{j=1}^{m\_2} k\_j=N} p(\mathbf{k},\beta)) A_0(N)]\right\} a\mathbf{e}$$

$$= z^{\wedge}\{\sum_{i=1}^{m\_2} n_i - 1\}\left\{zA_1 + z\sum_{N=1}^{\infty} z^N A_0(N) + (\sum_{i=1}^{m\_2} n_i h_i^0) A_2 + z\sum_{i=1}^{m\_2} n_i h_{i,i}\mathbf{I} + z\sum_{i=1}^{m\_2} n_i h_i^0(\mathbf{I}-\Gamma) + z\sum_{i=1}^{m\_2}\sum_{j=1:\ j\neq i}^{m\_2} n_i h_{i,j}\mathbf{I}\right\} a\mathbf{e}$$

$$= z^{\wedge}\{\sum_{i=1}^{m\_2} n_i - 1\}\left\{a\sum_{N=1}^{\infty} z(z^N - 1) A_0(N)\mathbf{e} + a\left[\sum_{i=1}^{m\_2} n_i (h_i^0 A_2 + zh_{i,i}\mathbf{I} + zh_i^0(\mathbf{I}-\Gamma) + z\sum_{j=1:\ j\neq i}^{m\_2} h_{i,j}\mathbf{I})\right]\mathbf{e}\right\}$$

$$= z^{\wedge}\{\sum_{i=1}^{m\_2} n_i - 1\}\left\{a\sum_{N=1}^{\infty} z(z^N - 1) A_0(N)\mathbf{e} + a(\sum_{i=1}^{m\_2} n_i h_i^0)(A_2 - z\Gamma)\mathbf{e}\right\}.$$

The last three equalities hold because the total sum of the probabilities of a multinominal distribution equals one, $[\Sigma_{N\geq 1}A_0(N)+A_1]\mathbf{e} = 0$, and $T\mathbf{e} + \mathbf{T}^0 = 0$, i.e., $\Sigma_j h_{i,j}+h_i^0 = 0$ for every i, respectively. Similarly, terms containing vectors $\{\mathbf{f_n}(2)\}$ on left hand side of inequality (5.1) can be evaluated. The result is given as

$$\sum_{i=1}^{m\_2} A(\mathbf{n},\mathbf{n}-\mathbf{e}_i)\mathbf{f}_{\mathbf{n}-\mathbf{e}\_i}(2) + A(\mathbf{n},\mathbf{n})\mathbf{f_n}(2) + \sum_{i=1}^{m\_2}\sum_{j=1:\ j\neq i}^{m\_2} A(\mathbf{n},\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j)\mathbf{f}_{\mathbf{n}-\mathbf{e}\_i+\mathbf{e}\_j}(2) + \sum_{\mathbf{k}\geq 0,\,\mathbf{k}\neq 0} A(\mathbf{n},\mathbf{n}+\mathbf{k})\mathbf{f}_{\mathbf{n}+\mathbf{k}}(2) \qquad (5.5)$$

$$= z^{\wedge}\{\sum_{i=1}^{m\_2} n_i - 1\}\left\{z\left[A_1 + \sum_{N=1}^{\infty} z^N A_0(N)\right]\mathbf{u} + (\sum_{i=1}^{m\_2} n_i h_i^0)(A_2 - z\Gamma)\mathbf{u}\right\}.$$

For $2\le j\le j_0$, terms containing vectors $\{\mathbf{f_n}(1, j)\}$ on the left hand side of inequality (5.1) are evaluated as follows.

$$\sum_{i=1}^{m\_2} A(\mathbf{n},\mathbf{n}-\mathbf{e}_i)\mathbf{f}_{\mathbf{n}-\mathbf{e}\_i}(1,j) + A(\mathbf{n},\mathbf{n})\mathbf{f_n}(1,j) + \sum_{i=1}^{m\_2}\sum_{t=1:\, t\neq i}^{m\_2} A(\mathbf{n},\mathbf{n}-\mathbf{e}_i+\mathbf{e}_t)\mathbf{f}_{\mathbf{n}-\mathbf{e}\_i+\mathbf{e}\_t}(1,j) + \sum_{\mathbf{k}\ge 0,\, \mathbf{k}\neq 0} A(\mathbf{n},\mathbf{n}+\mathbf{k})\mathbf{f}_{\mathbf{n}+\mathbf{k}}(1,j) \tag{5.6}$$

$$\begin{aligned}
&= a\Bigg\{\sum_{i=1}^{m_2} n_i h_i^0 A_2 z\wedge\{\sum_{t\in\Re(j)} n_t\} + \left[A_1 + \sum_{i=1}^{m_2} n_i h_{i,i}\mathbf{I} + \sum_{i=1}^{m_2} n_i h_i^0(\mathbf{I}-\Gamma)\right] z\wedge\{\sum_{t\in\Re(j)} n_t\} \\
&\quad + \left(\sum_{i=1:\, i\notin\Re(j)}^{m_2}\ \sum_{t=1:\, t\neq i,\, t\notin\aleph(j)}^{m_2} n_i h_{i,t}\mathbf{I}\right) z\wedge\{\sum_{t\in\Re(j)} n_t\} + \left(\sum_{i=1:\, i\notin\Re(j)}^{m_2}\ \sum_{t=1:\, t\neq i,\, t\in\Re(j)}^{m_2} n_i h_{i,t}\mathbf{I}\right) z\wedge\{1+\sum_{t\in\Re(j)} n_t\} \\
&\quad + \left(\sum_{i=1:\, i\in\Re(j)}^{m_2}\ \sum_{t=1:\, t\notin\Re(j)}^{m_2} n_i h_{i,t}\mathbf{I}\right) z\wedge\{\sum_{t\in\Re(j)} n_t - 1\} + \left(\sum_{i=1:\, i\in\Re(j)}^{m_2}\ \sum_{t=1:\, t\neq i,\, t\in\Re(j)}^{m_2} n_i h_{i,t}\mathbf{I}\right) z\wedge\{\sum_{t\in\Re(j)} n_t\} \\
&\quad + \sum_{\mathbf{k}:\ k_t=0,\ \text{for } t\in\Re(j)} \left[p(\mathbf{k},\beta)A_0(\sum_{i=1}^{m_2} k_i) z\wedge\{\sum_{t\in\Re(j)} n_t\}\right] \\
&\quad + \sum_{\mathbf{k}:\ \exists k_t>0,\ t\in\Re(j)} \left[p(\mathbf{k},\beta)A_0(\sum_{i=1}^{m_2} k_i) z\wedge\{\sum_{t\in\Re(j)} (n_t+k_t)\}\right]\Bigg\}\mathbf{e} \\
&= a z\wedge\{\sum_{t\in\Re(j)} n_t - 1\}\Bigg\{ z\sum_{\mathbf{k}:\ \exists k_t>0,\ t\in\Re(j)} \left[p(\mathbf{k},\beta)A_0(\sum_{i=1}^{m_2} k_i)(z\wedge\{\sum_{t\in\Re(j)} k_t\}-1)\right]\mathbf{e} \\
&\qquad + z(z-1)\sum_{i\notin\Re(j)} n_i(\sum_{t\in\Re(j)} h_{i,t})\mathbf{e} - (z-1)\sum_{i\in\Re(j)} n_i(\sum_{t\notin\Re(j)} h_{i,t})\mathbf{e}\Bigg\},
\end{aligned}$$

in which $[\Sigma_{N\ge 1}A_0(N)+A_1]\mathbf{e} = 0$, $T\mathbf{e} + \mathbf{T}^0 = 0$, i.e., $\Sigma_j h_{i,j}+h_i^0 = 0$, $A_2\mathbf{e}=\Gamma\mathbf{e}$ are used. The notation "$\exists$" stands for "there exists". Notice that $h_i^0=0$ for $i\in\Re(j)$ if $j\ge 2$. Summing up equations (5.4), (5.5), and (5.6) from $j=2$ to $j=j_0$, the left hand side of inequality (5.1) becomes

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

Rearranging terms in the third and fourth lines in equation (5.7) with respect to the subsets $\{\Im(j),\ 1\le j\le j_0\}$, inequality (5.1) becomes

$$
\begin{aligned}
z^{\{\sum_{i=1}^{m\_2} n_i -1\}}\Bigg\{ & a\sum_{N=1}^{\infty} z(z^N-1)A_0(N)\mathbf{e} - a(z-1)\Big(\sum_{i=1}^{m\_2} n_i h_i^0\Big)\Gamma\mathbf{e} \\
& + z\Big[A_1 + z\sum_{N=1} z^N A_0(N)\Big]\mathbf{u} + \Big(\sum_{i=1}^{m\_2} n_i h_i^0\Big)(A_2 - z\Gamma)\mathbf{u}\Bigg\} \qquad (5.8) \\
& + a(z-1)\sum_{i\in\Im(1)} n_i\Bigg[z\sum_{t=2}^{j\_0} z^{\{\sum_{l\in\Re(t)} n_l -1\}}\Big(\sum_{l\in\Re(t)} h_{i,l}\Big)\Bigg]\mathbf{e} \\
& + a\sum_{j=2}^{j\_0}\Bigg\{ z^{\{\sum_{i\in\Re(j)} n_i\}} \sum_{\mathbf{k}:\,\exists k\_t>0,\, t\in\Re(j)} \Bigg[p(\mathbf{k},\beta)A_0\Big(\sum_{i=1}^{m\_2} k_i\Big)\Big(z^{\{\sum_{i\in\Re(j)} k_i\}}-1\Big)\Bigg]\mathbf{e} \\
& + (z-1)\sum_{i\in\Im(j)} n_i\Bigg[\sum_{t=j+1}^{j\_0} z^{\{\sum_{l\in\Re(t)} n_l\}}\Big(\sum_{l\in\Re(t)} h_{i,l}\Big) - z^{\{\sum_{l\in\Re(j)} n_l -1\}}\sum_{t\in\Im(j-1)} h_{i,t}\Bigg]\Bigg\} \le -\varepsilon\mathbf{e}.
\end{aligned}
$$

Note that $\sum_{i\in\Re(j)} n_i\Big(\sum_{t\notin\Re(j)} h_{i,t}\Big) = \sum_{i\in\Im(j)} n_i\Big(\sum_{t\in\Im(j-1)} h_{i,t}\Big)$ by definition (equation (5.3)).

We now choose parameters z, $\mathbf{u}$, and a so that inequality (5.8), or equivalently inequality (5.1), holds for all but a finite number of $\mathbf{n}\in\aleph$ 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 $0<z<z_0$, $\sum_{n=1}^{\infty} z^n D_{n+t}$ is finite and uniformly bounded by $D^*(z)$-D_0 for all $t\ge 0$. 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 $z^{\{n_1+\ldots+n\_\{m_2\}-1\}}$, 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 $z^{\{n_1+\ldots+n\_\{m_2\}-1\}}$. In order to make inequality (5.8) true for all but a finite number of $\mathbf{n}\in\aleph$ 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 $(A_2\text{-}z\Gamma)\mathbf{u} = 0$ 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, 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 $a\sum_{N=1}^{\infty} z(z^N - 1)A_0(N)\mathbf{e}$ are $a\sum_{N=1}^{\infty} z(z^N - 1)(D_N \otimes \mathbf{I}\_\{m_1^s\})\mathbf{e}$ and $D^*(z)$ is finite, a small a can be chosen so that the last mm_1^s elements of $(A_1 + \sum_{N=1}^{\infty} z^N A_0(N))\mathbf{u}$ + $a\sum_{N=1}^{\infty} z(z^N - 1)A_0(N)\mathbf{e}$ 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 which is close to 1 and its corresponding vector $\mathbf{u}$. 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+…+n_{m_2}-1}, are less than -ε for some positive ε if at least one value in $\{n_i:\ i\in\Im(1)\}$ is large enough. Therefore, all elements of the sum of the first and second lines of inequality (5.8) are less than -εz^(n_1+…+n_{m_2}-1) for some positive ε if at least one value in $\{n_i:\ i\in\Im(1)\}$ 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 (5.1) holds for all but a finite number of $\mathbf{n}\in\aleph$ 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 $\{\mathbf{f_n}(1, j)\}$, $2\le j\le j_0$, are included in the test function to deal with the case when none in $\{n_i:\ i\in\Im(1)\}$ is large and some value in $\{n_i:\ i\notin\Im(1)\}$ 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 $\{n_i:\ i\in\Im(1)\}$ is large, the sum of the first, second, and third lines of inequality (5.8) is less than

$$
\begin{aligned}
& z\wedge\{\sum_{i=1}^{m\_2} n_i - 1\}\left\{-\varepsilon\mathbf{e} + a(z-1)\sum_{i\in\Im(1)} n_i\left[z\sum_{t=2}^{j\_0} z\wedge\{-\sum_{l\notin\Re(t)} n_l\}(\sum_{l\in\Re(t)} h_{i,l})\right]\mathbf{e}\right\} \\
& \le z\wedge\{\sum_{i=1}^{m\_2} n_i - 1\}\left\{-\varepsilon\mathbf{e} + az(z-1)\sum_{i\in\Im(1)} \frac{n_i}{z\wedge\{\sum_{l\in\Im(1)} n_l\}}\left[\sum_{t=2}^{j\_0}(\sum_{l\in\Re(t)} h_{i,l})\right]\mathbf{e}\right\} \\
& \le z\wedge\{\sum_{i=1}^{m\_2} n_i - 1\}\left\{-\varepsilon\mathbf{e} + \sum_{i\in\Im(1)} \frac{az^2(z-1)n_i}{\left[1 + 0.5(\sum_{l\in\Im(1)} n_l)^2(z-1)^2\right]}\left[\sum_{t=2}^{j\_0}(\sum_{l\in\Re(t)} h_{i,l})\right]\mathbf{e}\right\} \\
& \le -z\wedge\{\sum_{i=1}^{m\_2} n_i - 1\}(\varepsilon - \delta)\mathbf{e},
\end{aligned}
\tag{5.9}
$$

where δ can be arbitrarily small. Notice that $\Im(1)\subseteq\{i: i\notin\Re(j)\}$ when $j>1$ and $z^{n+1} = (1+(z-1))^{n+1} \geq 1+ n(n+1)(z-1)^2/2 \geq 1+ 0.5n^2(z-1)^2$. Thus, the sum of the first, second, and third lines of inequality (5.8) is less than $-\varepsilon z$^$(n_1+\ldots+n\_\{m_2\}-1)$ for some positive ε if at least one value in $\{n_i: i\in\Im(1)\}$ is large enough. (Note, to simplify the notation, we replace ε-δ by ε in the last line of inequality (5.9). Similar substitution shall take place in inequality (5.11) and in the discussion after equation (5.12).)

The last two lines of the left hand side of inequality (5.8) can be rewritten as

$$\begin{aligned}
&a\sum_{j=2}^{j_0} z\text{\^{}}\{\sum_{i\in\Re(j)} n_i -1\}\left\{ z \sum_{\mathbf{k}:\exists k_t>0,\, t\in\Re(j)} \left[p(\mathbf{k},\beta)A_0(\sum_{i=1}^{m_2} k_i)(z\text{\^{}}\{\sum_{i\in\Re(j)} k_i\}-1)\right]\right. \\
&\qquad\left. +(z-1)\sum_{i\in\Im(j)} n_i \left[z\sum_{t=j+1}^{j_0} z\text{\^{}}\{-\sum_{l\notin\Re(t),l\in\Re(j)} n_l\}(\sum_{l\in\Re(t)} h_{i,l}) - \sum_{t\in\Im(j-1)} h_{i,t}\right]\right\}\mathbf{e} \qquad (5.10)\\
&\leq \sum_{j=2}^{j_0} z\text{\^{}}\{\sum_{i\in\Re(j)} n_i -1\} a\left\{ z\sum_{N=1}^{\infty}(z^N-1)A_0(N) + \sum_{i\in\Im(j)} \frac{(z-1)z\,n_i}{\left(z\text{\^{}}\{\sum_{l\in\Im(j)} n_l\}\right)}\left(\sum_{t=j+1}^{j_0}(\sum_{l\in\Re(t)} h_{i,l})\right)\right. \\
&\qquad\left. -(z-1)\sum_{i\in\Im(j)} n_i\left(\sum_{t\in\Im(j-1)} h_{i,t}\right)\right\}\mathbf{e} \\
&\equiv \sum_{j=2}^{j_0} z\text{\^{}}\{\sum_{i\in\Re(j)} n_i -1\}\Delta(j,\mathbf{n}).
\end{aligned}$$

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 $z>1$, $n_i z\text{\^{}}\{-\sum_{l\in\Im(j)} n_l\}$ is a bounded function for every $i\in\Im(j)$. Then, for $2\leq j\leq j_0$, $\Delta(j, \mathbf{n})$ becomes negative when at least one value in $\{n_i: i\in\Im(j)\}$ is large enough. It follows that vector $\Delta(j, \mathbf{n})$ is uniformly bounded from above with respect to $\mathbf{n}\in \aleph$ and j (for any fixed z).

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

$$\begin{aligned}
&z\text{\^{}}\{\sum_{i=1}^{m_2} n_i -1\}\left\{-\varepsilon\mathbf{e} + a\sum_{j=2}^{j_0}\frac{\Delta(j,\mathbf{n})}{z\text{\^{}}\{\sum_{i\notin\Re(j)} n_i\}}\right\} \qquad (5.11)\\
&\leq z\text{\^{}}\{\sum_{i=1}^{m_2} n_i -1\}\left\{-\varepsilon\mathbf{e} + \frac{a}{z\text{\^{}}\{\sum_{i\in\Im(1)} n_i\}}\sum_{j=2}^{j_0}\Delta(j,\mathbf{n})\right\}.
\end{aligned}$$

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

To complete the proof, we still need to show that (5.1) holds when none in $\{n_i: i\in\Im(1)\}$ is large and some value in $\{n_i: i\notin\Im(1)\}$ is large. Similar to inequality (5.10), we denote the sum of the first three lines of inequality (5.8) as z^(n_1+…+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

$$z\wedge\{\sum_{i\in\Re(2)} n_i - 1\}\left[z\wedge\{\sum_{i\in\Im(1)} n_i\}\Phi(\mathbf{n}) + \Delta(2,\mathbf{n}) + \sum_{j=3}^{j\_0} \frac{\Delta(j,\mathbf{n})}{z\wedge\{\sum_{i\in\Re(2), i\notin\Re(j)} n_i\}} \right]. \tag{5.12}$$

Since $z\wedge\{\sum_{i\in\Im(1)} n_i\}$ is uniformly bounded from above for $\mathbf{n}$ in $\aleph$-$\aleph(1)$ and $\Phi(\mathbf{n})$ is uniformly bounded from above for $\mathbf{n}$ in $\aleph$, function $z\wedge\{\sum_{i\in\Im(1)} n_i\}\Phi(\mathbf{n})$ is uniformly bounded from above for $\mathbf{n}$ in $\aleph$- $\aleph(1)$. Thus, expression (5.12) is less than $-z\wedge\{\sum_{i\in\Re(2)} n_i - 1\}\varepsilon\mathbf{e}$ for some positive ε for $\mathbf{n}$ in $\aleph(1)$ and at least one value in $\{n_i: i\in\Im(2)\}$ is large enough, regardless of the values in $\{n_i: i\notin\Im(2)\}$. Let $\aleph(2)$ be the subset of $\aleph$-$\aleph(1)$ such that the left hand side of inequaity (5.8) is less than $-z\wedge\{\sum_{i\in\Re(2)} n_i - 1\}\varepsilon\mathbf{e}$ for some positive ε if at least one value in $\{n_i: i\in\Im(2)\}$ is large enough, regardless of the values in $\{n_i: i\notin\Im(2)\}$. Similarly, for $3\le j\le j_0$, we can find $\aleph(j)$, a subset of $\aleph$-$\aleph(1)\cup\ldots\cup\aleph(j-1)$ such that for $\mathbf{n}$ in $\aleph(j)$ the left hand side of inequality (5.8) is less than $-z\wedge\{\sum_{i\in\Re(j)} n_i - 1\}\varepsilon\mathbf{e}$ for some positive ε if at least one value in $\{n_i: i\in\Im(j)\}$ is large enough, regardless of the values of $\{n_i: i\notin\Im(j)\}$. It is easy to see that for any $\mathbf{n}$ in $\aleph(1)\cup\ldots\cup\aleph(j_0)$, inequality (5.1) holds for some positive ε. Since $\aleph$-$\aleph(1)\cup\ldots\cup\aleph(j_0)$ has only a finite number of members, we have proved that inequality (5.1) holds for all but a finite number of $\mathbf{n}\in\aleph$ for some positive ε, provided that a positive vector $\mathbf{u}$ can be found. To see why $\aleph$-$\aleph(1)\cup\ldots\cup\aleph(j_0)$ has a finite number of members, consider the special case with $m_2=2$, $\Im(1)=\{1\}$ and $\Im(2)=\{2\}$.

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

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

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

When $\rho<1$, the *MAP/PH/s* queue is ergodic (Assmusen [1]), which meant that the corresponding quasi-birth-and-death (QBD) Markov process $\{q(t), I(t), I_i(t), 1 \le i \le \max\{s, q(t)\}\}$ 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 $B^*(z) = B_{s+1,s} + zB_{s+1,s+1} + z^2B_{s+1,s+2} + \ldots$, for $1<z<z_0$. Let $\mathbf{y}$ be the unique solution to equations $\mathbf{y}B^*(1) = \mathbf{y}(B_{s+1,s} + B_{s+1,s+1} + B_{s+1,s+2} + \ldots) = 0$ and $\mathbf{ye} = 1$. Vector $\mathbf{y}$ is positive since $B^*(1)$ is irreducible. It can be verified that

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

Then it is easy to verify that $\mathbf{y}B_{s+1,s}\mathbf{e} = s\mu$ and $\mathbf{y}(B_{s+1,s+2}+2B_{s+1,s+3}+\ldots)\mathbf{e} = \lambda$. Thus, $\rho<1$ implies that Neuts' condition $\mathbf{y}(B_{s+1,s+2}+2B_{s+1,s+3}+\ldots)\mathbf{e} = \lambda < s\mu = \mathbf{y}B_{s+1,s}\mathbf{e}$ is satisfied. Denote by $sp(B^*(z))$ the eigenvalue with the largest real part of $B^*(z)$. Then $sp(B^*(1)) = 0$. Similar to the proof of Lemma 1.3.3 in Neuts [17], it can be proven that the derivative of $sp(B^*(z))$ at $z=1$ is negative. Thus, $sp(B^*(z)) < 0$ for z close to 1 and $1<z<z_0$.

Choose z such that $z < z_0$, z is close to 1, and $sp(B^*(z)) < 0$. $B^*(z)$ is an irreducible M-matrix (see Gantmacher [11]). Choose $\mathbf{v}$ to be the right eigenvector corresponding to $sp(B^*(z))$ with the first element to be one. Then $\mathbf{v}$ is positive and satisfies $B^*(z)\mathbf{v} = sp(B^*(z))\mathbf{v}$, $\mathbf{v}>0$ and $v_1 = 1$ ($\mathbf{v} = (v_1, v_2, \ldots, v\_\{mm_1^s\})$). Based on the special structure of A_2, choose $W_{s+K} = \mathbf{I}$, and

$$W_i = \begin{cases} (\mathbf{I}\_\{mm_1^i\} \otimes \boldsymbol{\alpha})W_{i+1}/z, & 0 \le i \le s-1; \\ W_{i+1}/z, & s \le i \le s+K-1. \end{cases} \qquad (5.15)$$

It can be verified that every element of vector $\mathbf{u}$ is positive, $(A_2\text{-}z\Gamma)\mathbf{u} = 0$, the last mm_1^s elements of $(A_1 + \sum_{N=1}^{\infty} z^N A_0(N))\mathbf{u}$ are given by $sp(B^*(z))\mathbf{v}$. When $K>0$, $B^*(z)\mathbf{v} = (B_{s+1,s} + zB_{s+1,s+1} + z^2B_{s+1,s+2} + \ldots)\mathbf{v} = sp(B^*(z))\mathbf{v} < 0$. When $K=0$, $B^*(z)\mathbf{v} = (B_{s,s-1}(\mathbf{I}\otimes\boldsymbol{\alpha}) + zB_{s,s} + z^2B_{s,s+1}+\ldots)\mathbf{v} = (B_{s+1,s} + zB_{s+1,s+1} + z^2B_{s+1,s+2} + \ldots)\mathbf{v} = sp(B^*(z))\mathbf{v} < 0$. Thus, vector $\mathbf{u}$ obtained from equation (5.13) satisfies our needs. This completes the proof.

Intuitively, vectors $\{\mathbf{f_n}(1,1)+\mathbf{f_n}(2)\}$ are used to guarantee that inequality (5.1) holds when n_i is large for $i \in \Im(1)$. The difficult part is to make the last mm_1^s elements of the left hand side of inequality (5.1) negative, which is achieved by using the *MAP/PH/s* queue with $\rho<1$. Vectors $\{\mathbf{f_n}(1, j)\}$ are used to guarantee that inequality (5.1) holds when n_i is large for $i \in \Im(j)$ and $j>1$.

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 at the end of Section 7.

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 (non-persistent) 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 $1-p$, $0\le p\le 1$. 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 $b = (p_y)/(s_h) < 1$.

Proof. Introduce the Markov process $\{N_i(t),\ 1\le i\le m_2,\ q(t),\ I(t),\ I_j(t),\ 1 \le j \le \max\{s, q(t)\}\}$ 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

$$A(\mathbf{n},\mathbf{n}+\mathbf{k}) = \begin{pmatrix} 0 & \cdots & 0 & \sum_{N=\sum_{j=1}^{m\_2} k\_j}^{\infty} [p(N,\mathbf{k},\beta) D_{s+K+N} \otimes \underbrace{\alpha \otimes \cdots \otimes \alpha}_{s}] \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & \sum_{N=\sum_{j=1}^{m\_2} k\_j}^{\infty} [p(N,\mathbf{k},\beta) D_{1+K+N} \otimes \mathbf{I}\_\{m_1^{s-1}\} \otimes \alpha] \\ 0 & \cdots & 0 & \sum_{N=\sum_{j=1}^{m\_2} k\_j}^{\infty} [p(N,\mathbf{k},\beta) D_{K+N} \otimes \mathbf{I}\_\{m_1^{s}\}] \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & \sum_{N=\sum_{j=1}^{m\_2} k\_j}^{\infty} [p(N,\mathbf{k},\beta) D_{K+N} \otimes \mathbf{I}\_\{m_1^{s}\}] \end{pmatrix} \tag{6.1}$$

$$\equiv \sum_{N=\sum_{j=1}^{m\_2} k_j}^{\infty} [p(N,\mathbf{k},\beta) A_0(N)],$$

where

$$p(N,\mathbf{k},\beta) = \frac{N!}{(N - \sum_{j=1}^{m\_2} k_j)!k_1!\cdots k_{m\_2}!} p^{\sum_{j=1}^{m\_2} k_j} (1-p)^{N-\sum_{j=1}^{m\_2} k_j} \beta_1^{k\_1}\cdots\beta_{m\_2}^{k\_m_2};. \qquad (6.2)$$

$$B_{i,j} = \begin{cases} \sum_{N=j-i}^{\infty} \binom{N}{j-i} p^{j-i}(1-p)^{N-j+i}(D_N \otimes \mathbf{I}\_\{m_1^i\}) \otimes \underbrace{\alpha \otimes \cdots \otimes \alpha}_{j-i}, & 0 \le i < j \le s; \\ \sum_{N=j-i}^{\infty} \binom{N}{j-i} p^{j-i}(1-p)^{N-j+i}(D_N \otimes \mathbf{I}\_\{m_1^i\}) \otimes \underbrace{\alpha \otimes \cdots \otimes \alpha}_{s-i}, & 0 \le i < s \le j \le s+K; \\ \sum_{N=j-i}^{\infty} \binom{N}{j-i} p^{j-i}(1-p)^{N-j+i}(D_N \otimes \mathbf{I}\_\{m_1^s\}), & 0 \le i < j \le s+K. \end{cases}$$

and

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

The necessity of the condition for ergodicity can be proven 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 (5.1). Inequality (5.8) then becomes

$$\begin{aligned} & z\wedge\{\sum_{i=1}^{m\_2} n_i - 1\}\Big\{ a\sum_{N=1}^{\infty} z[(zp+1-p)^N - 1]A_0(N)\mathbf{e} - a(z-1)(\sum_{i=1}^{m\_2} n_i h_i^0)\Gamma\mathbf{e} \\ & \qquad + z\left[A_1 + \sum_{N=1}(zp+1-p)^N A_0(N)\right]\mathbf{u} + (\sum_{i=1}^{m\_2} n_i h_i^0)(A_2 - z\Gamma)\mathbf{u}\Big\} \qquad (6.3) \\ & + a(z-1)\sum_{i\in\Im(1)} n_i \left[\sum_{t=2}^{j\_0} z\wedge\{\sum_{l\in\Re(t)} n_t\}(\sum_{l\in\Re(l)} h_{i,l})\right]\mathbf{e} \\ & + a\sum_{j=2}^{j\_0}\Big\{ z\wedge\{\sum_{t\in\Re(j)} n_t\} \sum_{\mathbf{k}:\exists k\_l>0,\, l\in\Re(j)} (z\wedge\{\sum_{t\in\Re(j)} k_t\}-1)\left[\sum_{N=\sum_{l=1}^{m\_2} k\_i}^{\infty} p(N,\mathbf{k},\beta)A_0(N)\right]\mathbf{e} \\ & \qquad + \sum_{i\in\Im(j)} n_i \left[\sum_{l=j+1}^{j\_0} z\wedge\{\sum_{t\in\Re(j)} n_t\}(\sum_{t\in\Re(l)} h_{i,t}) - z\wedge\{\sum_{t\in\Re(j)} n_t - 1\}\sum_{t\in\Im(j-1)} h_{i,t}\right]\mathbf{e}\Big\} \le -\varepsilon\mathbf{e}. \end{aligned}$$

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 of 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 goes orbiting again with probability q and leaves the queueing system with probability $1-q$, $0 \le q \le 1$.

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 $\{N_i(t), 1 \le i \le m_2, q(t), I(t), I_j(t), 1 \le j \le \max\{s, q(t)\}\}$ 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 (5.1), inequality (5.8) becomes

$$
\begin{aligned}
& z\wedge\{\sum_{i=1}^{m\_2} n_i - 1\}\Big\{ a\sum_{N=1}^{\infty} z[(zp+1-p)^N - 1]A_0(N)\mathbf{e} \\
& - a(z-1)(\sum_{i=1}^{m\_2} n_i h_i^0)[\Gamma + (1-q)(\mathbf{I}-\Gamma)]\mathbf{e} \\
& + z[A_1 + z\sum_{N=1}(zp+1-p)^N A_0(N)]\mathbf{u} + (\sum_{i=1}^{m\_2} n_i h_i^0)[A_2 + (1-z)(1-q)(\mathbf{I}-\Gamma) - z\Gamma]\mathbf{u}\Big\} \\
& + a\sum_{i\in\Im(1)} n_i \left[\sum_{l=2}^{m\_2} z\wedge\{\sum_{t\in\Re(l)} n_t\}(\sum_{k\in\Re(l)} h_{i,k})\right]\mathbf{e} \\
& + a\sum_{j=2}^{m\_2} z\wedge\{\sum_{l\in\Re(j)} n_l\}\Big\{ \sum_{\mathbf{k}:\exists k\_l>0,\, l\in\Re(j)} (z\wedge\{\sum_{l\in\Re(j)} k_l\} - 1)[\sum_{N=\sum_{l=1}^{m\_2} k\_i}^{\infty} p(N,\mathbf{k},\beta)A_0(N)]\mathbf{e} \\
& \qquad + \sum_{i\in\Im(j)} n_i \left[\sum_{l=j+1}^{m\_2} z\wedge\{\sum_{t\in\Re(j)} n_t\}(\sum_{k\in\Re(l)} h_{i,k}) - z\wedge\{\sum_{t\in\Re(j)} k_t - 1\}\sum_{k\in\Im(j-1)} h_{i,k}\right]\mathbf{e}\Big\} \le -\varepsilon\mathbf{e}.
\end{aligned}
\tag{6.5}
$$

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 $\mathbf{n} \in \aleph$ for some positive ε. For instance, choose z to be close to one and $\mathbf{u} = 0$. Other parameters can be determined accordingly. Notice that the test function of this case is simpler, since vectors $\{\mathbf{f_n}(2), \mathbf{n} \in \aleph\}$ 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 Diamond [3], Diamond and Alfa [4] to [7], and Neuts and Rao [19]). 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 enters the queue (or the server) immediately. 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.

The upper-bound queue. For a fixed positive integer N, define a retrial queueing similar to the $BMAP/PH/s/s+K$ retrial queue with PH-retrial times except that at most N customers in the orbit are trying to get service at any time. If a customer finds that there are N customers orbiting, the customer enters the orbit but do not start orbiting until the number of orbiting customers becomes less than N. For customer in the orbit waiting for orbiting, getting into orbiting follows a first-in-first-orbiting 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, or the orbit.

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

Proof. The necessity of $\rho<1$ can be proven 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.

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 proven by the sample path method used in Theorem 2. To prove sufficiency, consider the Markov process $\{N_i(t), 0\le i\le m_2, q(t), I(t), I_j(t), 1 \le j \le \max\{s, q(t)\}\}$ introduced in Section 3 except that $N_0(t)$ is introduced to record the number of customers who are in the orbit but not orbiting and $N_1(t)+\ldots+ N_{\{m\_2\}}(t) \le N$. $N_0(t)$ is positive only when $N_1(t)+\ldots+ N_{\{m\_2\}}(t) = N$. In the proof to Theorem 3, when one value in $\{n_i\}$ ($\mathbf{n} = (n_1, \ldots, n\_\{m_2\})$) is large enough, inequality (5.1) holds. Suppose that inequality (5.1) holds for $n_i>n_i^*$ for each i ($1\le i\le m_2$). 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 $\{N_i(t), 0\le i\le m_2, q(t), I(t), I_j(t), 1 \le j \le \max\{s, q(t)\}\}$ and inequality (5.1) holds for all but a finite states. In fact, according to the proof to Theorem 3, inequality (5.1) holds whenever $N_0(t)>0$, since $N_0(t)>0$ implies that $N_1(t)+\ldots+ N_{\{m\_2\}}(t) = N$ 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.

Theorem 1, Theorem 6, and Theorem 7 show that the *MAP/PH/s/s+K* retrial queue with *PH*-retrial times, the two modifications (lower-bound and upper-bound queues), and the classical *MAP/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 an orbiting customer. 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 orbiting. 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 orbiting customers is not large. Fortunately, the loss of capacity is recovered when the number of orbiting customers 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 $\{N_1(t), q(t), I(t), I_j(t), 1 \le j \le \max\{s, q(t)\}\}$ is a quasi-birth-and-death Markov process, and 2) it has the *M/M/s/s* retrial queue with exponential retrial times, and *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 solution.

References

[1] Asmussen, S., Applied Probability and Queues, John Wiley & Sons, Chichester New York, 1987.

[2] Cohen, J.W., *The Single Server Queues*, North Holland, Amsterdam, 1982.

[3] Diamond, J., *Matrix analytic methods for retrial queues*, Ph.D Thesis, Department of Mechanical and Industrial Engineering, University of Manitoba, 1995.

[4] Diamond, J. and A.S. Alfa, Matrix analytic methods for *M/PH/1* retrial queues, *Stochastic Models*, **11**, 447-470, 1995.

[5] Diamond, J. and A.S. Alfa, The *MAP/PH/1* retrial queue, *Stochastic Models*, **14**, 1998 (to appear)

[6] Diamond, J.E. and A.S. Alfa, Matrix analytical methods for retrial queues with finite buffers, 1997 (submitted to QUESTA)

[7] Diamond, J.E. and A.S. Alfa, Matrix analytical method for multiserver retrial queues with finite buffers, accepted for presentation at the Workshop on Retrial Queues (WRQ98), Madrid, September 1998.

[8] Falin, G.I., On sufficient conditions for ergodicity of multi-channel queueing systems with repeated calls, *Adv. Appl. Prob.*, **16**, 447-448, 1984.

[9] Falin, D., A survey of retrial queues, *Queueing Systems*, **7**, 127-168, 1990.

[10] Falin, G.I. and J.G.C. Templeton, *Retrial Queues*, Chapman & Hall, 1997.

[11] Gantmacher, F.R., *The Theory of Matrices*, Chelsea, New York, 1959.

[12] Kulkarni, V.G. and H.M. Liang, Retrial queues revisited, *Frontiers in Queueing: Models and Applications in Science and Engineering*, Edited by Dshalalow, J.H., CRC Press, 19-34, 1997.

[13] Liang, H.M. and V.G. Kulkarni, Stability condition for a single server retrial queue, *Adv. Appl. Probab*. **25,** 690-701, 1993.

[14] Li, Hui and T. Yang, The steady-state distribution of the *PH/M/1* retrial queue, The Second International Conference on Matrix Analytic Methods, Winnipeg, Manitoba, Canada, 1998.

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

[16] Neuts, M.F., A versatile Markovian point process, *J. Appl. Prob.*, **Vol 16** (1979), 764-779.

[17] Neuts, M.F., *Matrix-Geometric Solutions in Stochastic Models: An algorithmic approach*, The Johns Hopkins University Press, Baltimore, 1981.

[18] Neuts, M.F., *Structured Stochastic Matrices of M/G/1 Type and Their Applications*, New York: Marcel Dekker Inc., 1989.

[19] Neuts, M.F. and B.M. Rao, Numerical investigation of a multiserver retrial model, *Queueing Systems*, **7**, 169-189, 1990.

[20] Yang, Tao and J.G.C. Templeton, A survey on retrial queues, *Queueing Systems* **2**, 201-233, 1987.

[21] Yang, Tao, M.J.M. Posner, J.G.C. Templeton, and Hui Li, An approximation method for the *M/G/1* retrial queue with general retrial times, *E.J.O.R.*, **76**, 552-562, 1994.